

→ 郭晓科 主编

这是一场革命，我们现在做的只是冰山一角，但是由于庞大的数据新来源而带来的定量化方法，将横扫学界、商界和政界，所有领域都将被触及。

——哈佛大学定量社会研究所主任加里·金（Gary King）

最新的、有重要价值的国际研究报告、国际数据、行业最新进展等文献，为读者展示了一幅浩瀚的大数据景观，无论对于专业人士，还是对于普通公众，面对汹涌来袭的大数据时代，这本书无疑具有重要价值。

——李希光

清華大學出版社

大数据
Big Data

郭晓科　主编

清华大学出版社
北京

内 容 简 介

大数据的广泛应用已经彻底地改变了人类世界，这场信息革命的号角正在吹响，本书集纳了全球关于"大数据"（Big Data）的最新研究成果，为读者清晰勾勒出一幅"大数据"在社会各领域被广泛应用的广阔图景，并前瞻性地描绘了未来的大数据世界。

本书既是一本科普读物，让读者了解什么是大数据、大数据的应用、大数据的未来，以及大数据的潜在商业价值；同时对于公共政策、信息科学、社会科学等领域的交叉融合具有启发意义。

图书在版编目（CIP）数据

大数据/郭晓科主编.--北京：清华大学出版社，2012
ISBN 978-7-302-30230-8

Ⅰ.①大…　Ⅱ.①郭…　Ⅲ.①数字技术—普及读物　Ⅳ.①TP3-49

中国版本图书馆 CIP 数据核字（2012）第 229692 号

责任编辑：纪海虹
封面设计：蒋　宏
责任校对：王荣静
责任印制：

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　**邮　　编：**100084
社 总 机：010-62770175　**邮　　购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：
装 订 者：
经　　销：全国新华书店
开　　本：170mm×240mm　**印　　张：**8.5　**字　　数：**154 千字
版　　次：2013 年 1 月第 1 版　**印　　次：**2013 年 1 月第1 次印刷
印　　数：1～5000
定　　价：　.00 元

产品编号：049668-01

序

随着“大数据”时代的到来，“数据”这种抽象的东西，在我的日常生活中变得越来越具体和重要。有一次去东三环燕莎附近的一家餐馆吃饭，当我回到停车场时，发现车的后备箱被人撬开，我的笔记本电脑被偷走了，虽然这台笔记本电脑价值 1 万多元，让我心疼的却是存在硬盘里的资料，它们的价值可能是电脑的几十倍、上百倍。这件事让我付出了惨痛的代价，也让我养成了一个好习惯，资料随时备份在移动硬盘、办公室电脑等多个终端里。

第一次接触电脑，还是 20 世纪 80 年代初我在中科院理论物理所给周光召所长做外事秘书时，周先生当时花了 250 美元从美国带来了一台刚面市的苹果电脑，让我第一次看到了电脑视窗，第一次意识到电脑里的数据可以看得见、看得懂，而且还可以用来玩游戏。而在此之前的 20 世纪 60 年代，周光召、于敏、何祚庥等科学家做核武器的理论设计，使用的还是手摇计算机。手摇计算机只能做简单的数学运算，例如加、减、乘、除、开根号、求平方等，如果需要输入三角函数和对数，都需要查表，使用起来也十分麻烦，经常需要正摇几圈，反摇几圈，还要用纸、笔记录。

研究生毕业以后，我被分配到新华社做记者，恰好赶上新华社全面采用王安电脑终端编辑英文新闻，使英文新闻的编发工作实现了电脑化。在此之前，记者要靠纸、笔、打字机、传真机和电传机进行手工作业，编辑部门要把编好的稿件送到发稿部门，由报务人员按稿件先打字作电传孔条，然后再在各条线路的发送机上发出。使用终端机后，编辑、记者可以在电脑上直接对稿件进行编辑修改。

我买的第一台个人电脑是中关村的组装机。那台电脑的操作系

统是 UCDOS，使用的是 5.25 寸的软盘，容量是 360K。我的第二台个人电脑的软盘已经更新成 3.5 寸，容量也增加到 1.44M，我的办公室现在还有上百张 3.5 寸的软盘，可惜现在已经不太容易找到能够读软盘的电脑了。光驱、移动硬盘、U 盘打败了软盘，成为数据存储的主流，存储介质的容量也越来越大，几年前的 U 盘一般都是 16M 的，而如今 16G 的 U 盘都嫌不够用。

我每次出国都要逛电子产品商店，看看有没有容量更大的移动硬盘或 U 盘。我现在使用的是两块 1T 的移动硬盘，分别存储不同的文件。其中一块移动硬盘里存储着 Foxmail 邮件客户端信息，数据容量已经达到 20G，存储着我所有的电子邮件，我用 Foxmail 对邮件信息进行了分类，它就像我的一个私人图书馆一样，随时可以在其中方便地查找资料。

但是，使用移动硬盘仍然不够方便和安全，因为一旦移动硬盘丢了、坏了，或者是当你急用的时候却发现它不在身边，都会带来不便。最好的方式是通过云储存，把资料备份到云端，但最大的担心是怕不安全，万一云端的服务器出了问题，或者网络出了问题，造成的损失是不可估量的，所以我现在还是靠自己储存数据。

我的办公室里有一面墙的书架都是存放录音带、录像带、光盘和各类软盘的，存储着我从教书以来的所有影像、课件、资料，但是查找文件就像大海捞针一样。现在许多软件对数据分类和查找都支持得很好，如果能在一张移动硬盘上集中存储这些数据，自然要方便很多。但移动硬盘也不安全，既有被盗的风险，也担心使用中被损坏。所以最终的解决方案还是要依靠云存储，所有的文件都存储在虚拟空间里，随时可以通过互联网找出来。

我对“网络数据”的直观感受就是它的增长速度像原子弹链式反应一样，至今仍在加速膨胀。我从 1995 年开始使用互联网，一直到 2010 年的 15 年时间里，积累的数据资料也就只有书架上的各类光盘、软盘、磁带，而从 2010 年到现在这 2 年的时间，就积累了 2T 的数据，可能与前 15 年数据的容量相当。

在大数据时代，人们对数据的分类检索和储存智能化要求越来越高，否则查找有用的数据就像沙里淘金一样，大数据对人们来说意味着宝藏，同时也不可避免地带来了数据垃圾。作为一名研究人员，我从事研究工作的基础是文献检索和综述，离不开数据的收集、分类、综述和摘要，这些工作在过去都是依靠纸质的报刊、图书文献，工作的方法是“剪报”。我从小学三年级开始剪报，一直到读研究生还保持着这个习惯。后来到新华社当记者，查找文献还是依靠

剪报。再后来出现了电脑，这的确给文献的收集和使用带来了方便，但由于太相信电脑，一旦系统崩溃或硬盘坏了，数据就没了。

美国皮尤中心于2012年7月20日发布了一项关于大数据的民意调查报告。有53%的人乐观地认为大数据的应用能够促进社会、政治和经济的智能化发展。但同样值得注意的是，39%的人对大数据的前景表现出极大的忧虑，认为大数据会造成人类对自身预测能力的盲目自信，进而会导致很多错误的决定，甚至有人认为大数据的崛起对于整个社会而言，无疑是一个噩梦。人类最终不是被核武器毁灭，而是被“堆积如山、臭气熏天”的互联网信息埋葬。

在我看来，大数据时代人类面临的更大问题是，各类互联网终端的普及，特别是社交媒体的出现，使得人人可以成为文章的作者，人人都是学者，人人都是杂文家，人人都是摄影家，人人都是制片人，人人都是演员。在纸媒时代，作者的门槛很高，那个时候在书店里随便买一本书都有价值，今天由于信息的爆炸式增长，人们很难区分什么是有价值的信息，什么是垃圾信息，灾难来自于把这两类信息都放在一个数据库里，有价值的信息被埋没，信息的泛滥带来人类整体的平庸化和低俗化。人生是短暂的，如果人生要在荒诞的信息中自娱自乐，一辈子就那样过去了。问题是人类意识不到自己是处在高雅文化传播当中还是低俗平庸文化传播当中。人类将来也许不是被核武器毁灭，而是被“大数据”所毁灭。因此，我们要研究和掌握其中的规律。媒体教育工作者要面向公众开展大数据的媒介素养教育，研究人员要开发出最好的技术，使人们能够利用适当的技术方便地获取更多有价值、有深度的信息。

由郭晓科博士主编的《大数据》，虽然篇幅不长，但本书编者们通过艰苦的文献研究，精心编辑了最新的、有重要价值的国际研究报告、重要数据、行业最新进展等文献，从数据大爆炸、大数据应用、大数据挖掘、大数据前瞻、大数据安全等不同的角度为读者展示了一幅浩瀚的大数据景观。面对汹涌来袭的大数据，无论对于专业人士还是普通公众，这本书无疑具有重要价值。

李希光

清华大学国际传播研究中心主任

联合国教科文组织媒介素养教席主任

2012年8月于清华园

FOREWORD 前言

20 年前，“数据”对于普通人来说，还是一个相当专业的词汇。时至今日，“数据”已经无孔不入地渗透到我们的生活。人们在日常生活和工作中收发邮件和短信、拍照、录像、撰写文稿、计算机绘图及编程，每天都在源源不断地产生大量的数据。全人类一年产生的数据量以及它的增长速度都大得惊人：全球著名咨询机构 IDC（国际文献资料中心）在 2006 年估计全世界产生数据量为 0.18ZB(1ZB=100 万 PB)，而截至 2011 年这个数字已经提升了一个数量级，达到 1.8ZB，相当于全世界每个人一块用 100 多 GB 的硬盘存储的数据。这种增长仍在加速，预计 2015 年将达到近 8ZB。

2011 年 6 月，麦肯锡全球研究所（MGI）发布了《大数据：创新、竞争和生产力的下一个前沿》(*Big data: The next frontier for innovation, competition, and productivity*)。在这份报告中，“大数据”的概念得到了清晰阐释，麦肯锡在研究报告中指出，数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产因素；而人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来。

2012 年 1 月在瑞士举行的达沃斯世界经济论坛发布了一份名为《大数据，大影响》(*Big Data, Big Impact*) 的报告，再次引起全球关注与热议。这份报告认为，大数据就像货币和黄金一样，是一种新型的经济资产。多家投资机构甚至据此判断，“大数据”将会成为贯穿 2012 年的一条全新投资主线。

哈佛大学定量社会研究所主任加里·金（Gary King）在接受《纽约时报》记者史蒂夫采访时说：“这是一场革命，我们现在做的只是

冰山一角，但是由于庞大的数据新来源而带来的定量化方法，将横扫学界、商界和政界，所有领域都将被触及。”

2012 年 3 月 29 日，美国联邦政府宣布了《大数据研究和发展倡议》（*Big Data Research and Development Initiative*），斥资 2 亿美元投入大数据研究领域，以加强政府各个部门、研究机构和其他组织从大量复杂的数据中提取、分析重要信息的能力。这一倡议涉及美国联邦政府的六个部门，分别是美国国家科学基金、美国国家卫生研究院、美国能源部、美国国防部、美国国防部高级研究计划局和美国地质勘探局。这些部门将大力推动和改善与大数据相关的收集、组织和分析工具及技术的研发和使用，力图在科学发现、环境保护和生物医药研究、教育、国家安全及战争策略等领域利用大数据能力取得突破。

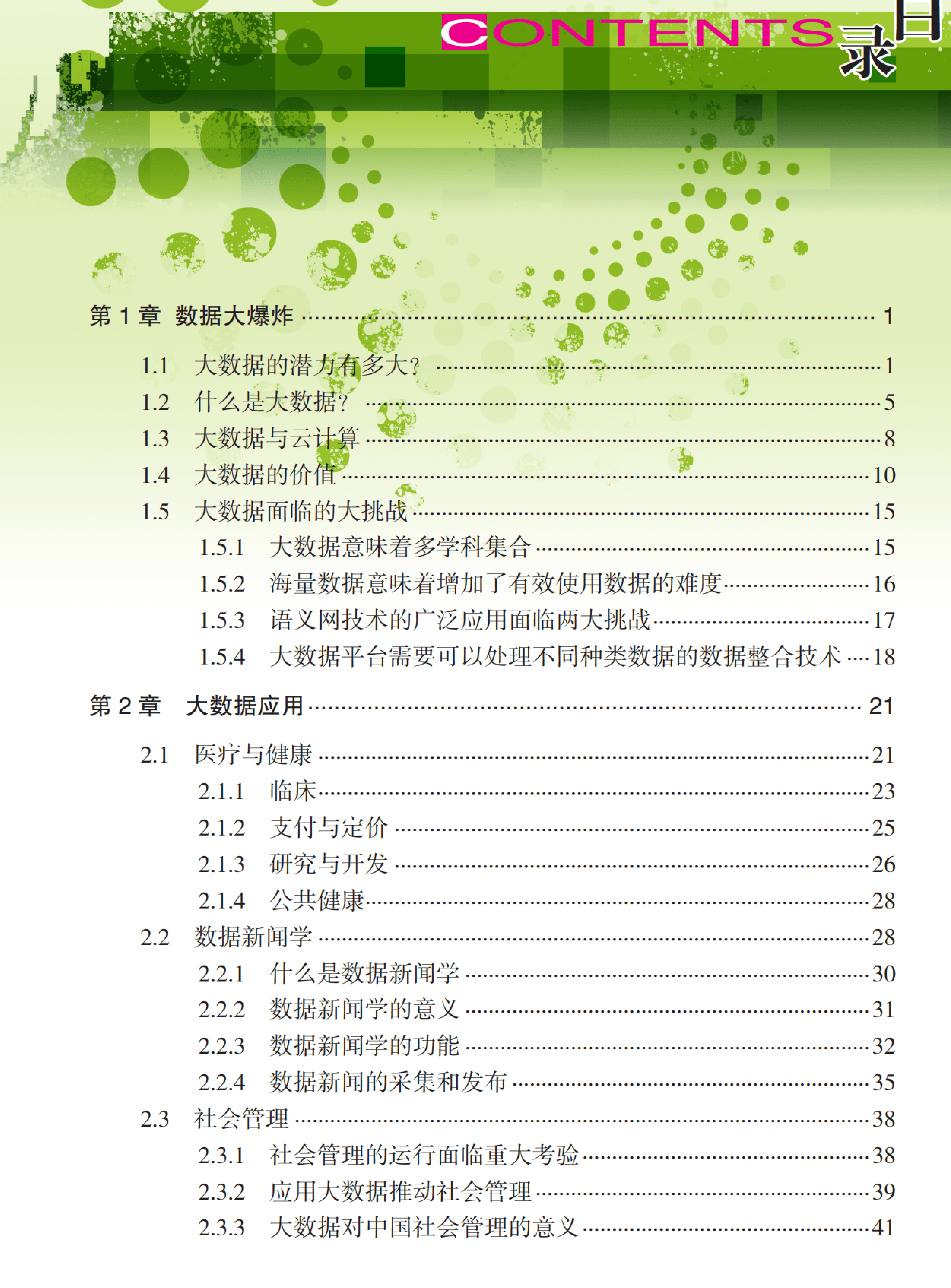
中国已经成为世界第二大经济体，但不容忽视的是高增长的代价十分高昂，从总体上来看，中国仍处于全球经济食物链的底端，用高能耗、高污染、廉价劳动力维系经济增长的模式难以为继，中国在蒸汽机革命和电气化革命中都落后于世界，但在“大数据时代”不能再落后。我们拥有前所未有的历史机遇：中国不仅拥有世界上最多的人口，到 2005 年底，中国的高等学校有 2300 余所，在学大学生总数已超过 2300 万人，高等教育总体规模已位居世界第一位。① 中国的大学培养了大量的 IT、数据统计、社会管理等专业人才，为中国的大数据战略进行了很好的人才储备。能否在“大数据”时代抓住历史机遇，成为全球信息革命的主角，是实现中国经济结构转型和中华民族伟大复兴的重要因素。

这本书旨在为中国政界、产业界、教育界以及社会各界人士打开一扇了解“大数据”的窗户，通过对麦肯锡全球研究所、国际文献资料中心、皮尤研究中心等全球著名咨询机构最新发布的有关大数据的报告进行编译，深入浅出地介绍了什么是大数据、大数据的价值、大数据的应用、大数据的挖掘、大数据的未来、大数据的安全等内容。在编译的过程中，清华大学国际传播研究中心的助理研究员刘娟、张小娅、汪震、刘沙沙、周燕各自负责一章，为了尽快把国际最新、最权威动态成果介绍给国内读者，她们付出了巨大的努力，在此表示感谢。

郭晓科

2012 年 7 月于清华园

① 上述这段话来自教育部前部长周济于 2006 年 7 月 13 日在第三届中外校长论坛的讲话。来源：http://news3.xinhuanet.com/newscenter/2006-07/13/content_4829159.htm。



CONTENTS 目录



数据大爆炸

欢迎来到“大数据时代”！

2012 年 1 月在瑞士举行的达沃斯世界经济论坛上，一份名为《大数据，大影响》（*Big Data, Big Impact*）的报告引起热议。这份报告认为，大数据就像货币和黄金一样，是一种新型的经济资产。多家投资机构甚至据此判断，“大数据”将会成为一条全新投资主线。

1.1 大数据的潜力有多大?

2012 年 2 月，科学记者史蒂夫•洛尔（Steve Lohr）① 在《纽约时报》撰文写道：科学、体育、广告、公共健康等各个不同领域，都越来越趋向基于数据的发现和决策。

“这是一场革命，我们现在做的只是冰山一角，但是由于庞大的数据新来源而带来的定量化方法，将横扫学界、商界和政界，所有领域都将被触及。”哈佛大学定量社会研究所主任加里 • 金（Gary King）在接受《纽约时报》记者史蒂夫采访时说。

麻省理工大学管理学院的经济学家埃里克 • 布吕诺尔夫松（Erik Brynjolfsson）接受采访时说，要充分理解大数据的潜在影响，必须通过显微镜来观察。四个世纪前发明的显微镜使人们前所未有地在细胞层面观察和测量事物。这是度量方面革命性的举措。数据测量就是现代人的显微镜。例如，谷歌（Google）的搜索、脸谱（Facebook）的帖子和推特（Twitter）的信息，使得细

① Steve Lohr，《纽约时报》记者，专长科技、商业和经济报道。1990 年前曾任《纽约时报》驻外记者、编辑。著有《创造软件革命的程序员们》（2001）一书。

节化、即时化的测量行为和情绪成为了可能。

布吕诺尔夫松说，在商业和经济等其他领域，决策将越来越依赖于数据和分析，而非经验和直觉。

有很多事件可以说明以数据为先的思考方式所带来的好处。最著名的还是 2003 年迈克尔 • 刘易斯（Michael Lewis）所写的著作《点球成金》（*Money ball*），讲述了低预算的奥克兰运动家队如何通过整理后的数据和晦涩的棒球统计资料来发现被低估的运动员。繁重的数据分析不仅在棒球领域成为一项标准，而且也被应用于其他运动，如英国足球远在去年由布拉德 • 皮特主演的电影版《点球成金》之前，就已经使用了数据分析。

沃尔玛、科尔之类的零售商也通过分析销售情况、定价，以及经济、人口统计、天气等数据，决定不同店铺的产品类别及降价促销的时机。

全球国际快递（UPS）等运输公司通过采集卡车运输次数和路线来调整运输的路线。

Match.com 等在线相亲服务网站，经常查看在线用户的个人情况、反应情况和沟通情况，以提高男女约会安排的匹配率。

美国的警察部门，以纽约为首，使用了计算机化的人像绘图，并通过分析逮捕记录、发工资日、体育赛事、降雨和节日等变量来试图预测可能发生犯罪的“热点区域”，从而可以提前在这些地区部署警力。

布吕诺尔夫松和其他两名同事 2011 年发表的研究表明，由数据引导的管理方式正在美国企业界蔓延，而且开始出现回报。他们研究了 179 家大型企业，发现那些正在使用“基于数据的决策方式”的企业所获得的利润比其他企业高出 5 至 6 个百分点。

大数据的预测能力正在被发掘，并在公共健康、经济发展和预测等领域显示出了成功的希望。研究人员发现，就在某一地区医院的急诊室里流感病人增加的约两周前，谷歌搜索里对“流感症状”和“流感治疗”的搜索请求出现了一个小高峰（急诊室的报告通常比实际接诊情况晚两周左右）。

据国际数据资讯公司（Global Pulse）估测，数据数量一直在增加，每年增长 50%。这个速度不仅是指数据流的增长，而且还包括全新的数据种类的增多。如今全球有数不清的数据感应器，应用于工业设备、汽车、电子量表、集装箱等。它们可以测量并传递地点、移动、振动、温度、湿度，甚至空气中的化学变化。将这些传递沟通的感应器与计算机智能连接起来，你就能看到物联网和工业互联网的崛起。

如果说数据是新型经济资产，那么赋予数据生产力的则是互联网。

2011 年 5 月，全球知名咨询公司麦肯锡（McKinsey & Company）的研究

部门麦肯锡全球研究所（McKinsey Global Institute）在对13个国家（法国、美国、英国、德国、日本、意大利、加拿大、俄罗斯、瑞典、韩国、中国、巴西、印度）调研后发布了题为《互联网的价值：网络对经济增长、就业及繁荣的影响》（*Internet matters: the net's sweeping impact on growth, jobs, and prosperity*）的调查报告。报告从7个方面论述了互联网对经济增长的贡献（Matthieu et al., 2011）。

（1）互联网体量巨大，发展迅猛。全球网民数量已达20亿，并仍在以每年2000万的速度增长。互联网经济平均已占13个国家GDP的3.4%，超过农业和能源，成为经济的巨大推动力。

（2）各国互联网经济发展水平差异大，在瑞典和英国这样的发达国家，互联网经济占GDP的比重是6%，而在13个国家中有9个，这一数字仍在4%以下。全球互联网经济的发展潜力不容小觑。

（3）互联网经济对GDP贡献巨大。报告中称，从1995年到2009年的15年间，数据覆盖的13个国家互联网经济占GDP增长的7%，并且影响在持续扩大；在互联网产业发达的国家（瑞典、德国、英国、法国、美国、韩国、加拿大、意大利、日本），15年间互联网对GDP增长的贡献率达10%，15年间的最后5年（2005—2009年），这一数字翻了一番，达21%；即使在互联网产业发展中国家（中国、印度、巴西、俄罗斯），互联网经济对GDP增长的贡献也达到了3%（见图1-1）。

（4）互联网产业的发达程度与提升生活质量关系紧密。相关研究显示，过去的15年中，互联网发展给被调查的13个国家所带来的人均GDP增长约为500美金，相当于19世纪工业革命历经50年取得的成绩。

（5）相较于传统行业，互联网更能催生就业机会。法国的一项研究显示，过去的15年，互联网每摧毁1个就业岗位，便新创造2.4个就业岗位。麦肯锡全球调查数据支持了这一论断，并把这一数字更新为2.6个。

（6）互联网促使经济发展现代化。互联网的使用为中小企业提升了10%的生产力。倚力于互联网技术的中小企业，增长速度和产品出口量都是同类企业的2倍。

（7）互联网创造令人惊异的消费者剩余[①]。月均每用户的消费者剩余从德国的13欧元到英国的20欧元（见图1-2）。仅2009年一年，法国创造的消费者剩余为70亿欧元，美国为460亿欧元。

---

① 消费者剩余（consumer surplus），经济学概念，指消费者为获得一种商品所愿意支付的价格与他实际支付的价格之间的差额。消费者网上消费所产生的剩余除了价格差额之外，还要减去网络交易成本，以及因广告等因素可能产生的任何形式的污染成本。

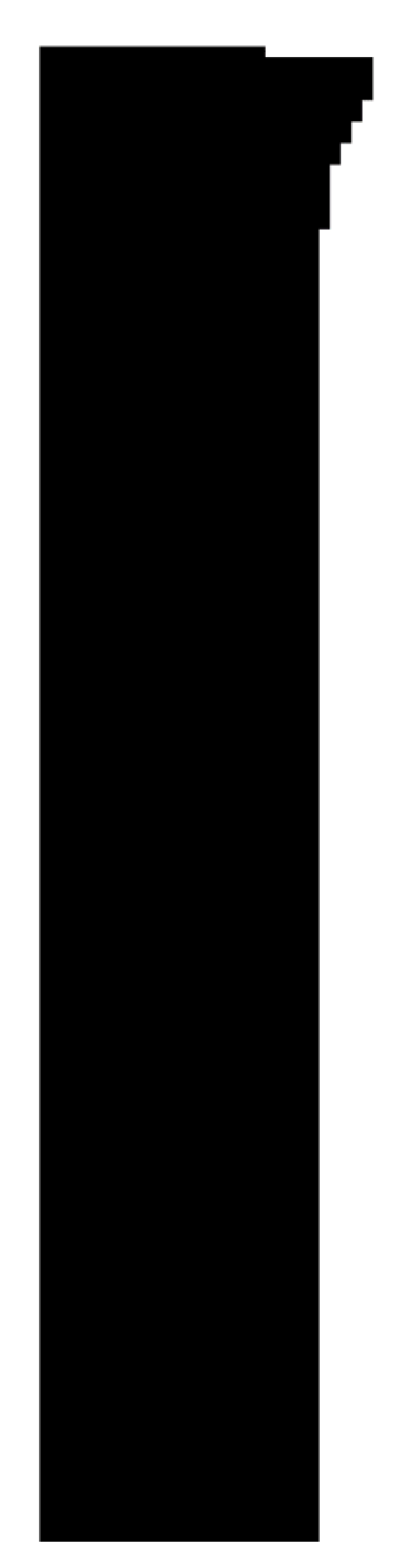


图1-1 互联网对13个国家GDP增长的贡献[①]



图1-2 各国互联网用户产生的消费者剩余价值[②]

① 资料翻译自：麦肯锡全球研究所报告《互联网的价值：网络对经济增长、就业及繁荣的影响》。
② 资料翻译自：麦肯锡全球研究所报告《互联网的价值：网络对经济增长、就业及繁荣的影响》。

在同一份报告中，麦肯锡全球研究所指出使用互联网或物联网收集市场和消费者信息的公司，很快会发现它们的数据库已被浩瀚的信息充塞。传统的数据处理工具已无法实现数据收集、储存、检索、共享、分析和视觉化的功能。倘若不加整理，互联网终有一天会成为克利夫•斯托尔（Cliff Stoll）[①]曾担心的“塞满垃圾信息的荒野”——“大数据”的意义正在于寻找管理海量数据的科学路径。报告指出长于“大数据时代”的人士将迎来各种各样的机遇。《纽约时报》的分析文章认为美国至少还需要14万至19万具备“深厚分析”专业技能的人才，以及150万熟悉数据的经理级人才（Steve, 2012）。麦肯锡预计下一个10年美国健康医疗业的年均增长约1%，亦即超过3000亿美元的潜在商业价值；欧洲发达国家的公共事业年均增长约为0.5%，也就是2.55千亿欧元的市场。

在这份预测大数据带来大市场的报告发布一个月之后，2011年6月，麦肯锡全球研究所又密集地发布了另一份报告：《大数据：创新、竞争和生产力的下一个前沿》（*Big data: The next frontier for innovation, competition, and productivity*），在这份报告中，“大数据”概念得到了清晰阐释。

1.2　什么是大数据？

“大数据”一词由英文“Big Data”翻译而来，过去常说的“信息爆炸”、“海量数据”等已经不足以描述这个新事物。麦肯锡全球研究所报告《大数据：创新、竞争和生产力的下一个前沿》对“大数据”定义如下（James, 2011）：

大数据是指大小超出了传统数据库软件工具的抓取、存储、管理和分析能力的数据群。这个定义有意地带有主观性，对于“究竟多大才算是大数据”，其标准是可以调整的，即：我们不以超过多少TB（1000GB）为大数据的标准。我们假设随着时间的推移和技术的进步，大数据的“量”仍会增加。还应注意到，该定义可以因部门的不同而有所差异，这取决于什么类型的软件工具是通用的，以及某个特定行业的数据集通常的大小。因此，今天众多行业的大数据范围可以从几十TB到数千TB。

作为特指的大数据，按EMC[②]的界定，其中的“大”是指大型数据集，一般在10TB规模左右；多用户把多个数据集放在一起，形成PB级的数据量；

---

① Cliff Stoll，美国天文学家、作家。亚利桑那大学行星科学博士。著有《杜鹃蛋》（*The Cuckoo's Egg*，1989），《硅谷蛇油》（*Silicon Snake Oil - Second thoughts on the information highway*,1995），《高科技异端者》（*High-Tech Heretic: Reflections of a Computer Contrarian*, 2000）。

② EMC公司，1979年成立于美国马萨诸塞州霍普金市，1989年开始进入企业数据储存市场。提供信息存储及管理产品、服务和解决方案。截至2011年，EMC在中国的北京、上海、广州等地设立了16家分公司。

同时这些数据来自多种数据源，以实时、迭代的方式来实现。大数据通常与Hadoop、NoSQL、数据分析与挖掘、数据仓库、商业智能以及开源云计算架构等诸多热点话题联系在一起。

IBM 负责软件和硬件两大集团的高级副总裁 Steve Mills 在 IBM 2011 IOD 大会上说："分析不再是一个工具，而是一项必要的能力，能让业务流程智慧运转的能力。企业必须将对信息的洞察力转化为行动，不是为了获得竞争优势，而是因为它已经变成生存的根本。"（见图 1-3）

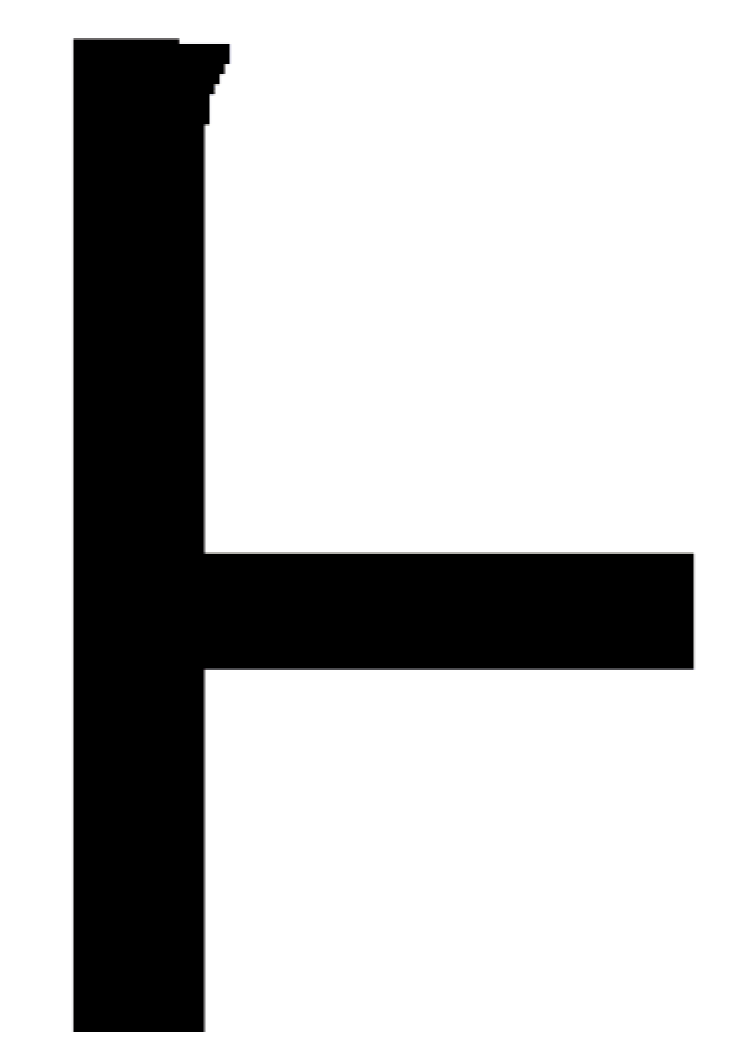


图1-3 IBM高级副总裁Steve Mills在IOD大会上详述IBM的大数据策略①

IBM 公司把大数据概括成了三个V，即大量化（Volume）、多样化（Variety）和快速化（Velocity），并向客户推出了"大数据解决方案"服务。IBM 公司所概括的这三个大数据的特点也反映了大数据所潜藏的价值（Value），或许可以认为，这四个V就是大数据的基本特征。

"大数据"的首要特征是数据量大。基于电脑的数据储存和运算是以字节（byte）为单位的，1KB（Kilobyte）=1024B，又称千字节；更高级的数量单位分别是 1MB（Megabyte，兆字节）、1GB（Gigabyte，吉字节）、1TB（Trillionbyte，太字节）、1PB（Petabyte，拍字节）、1EB（Exabyte，艾字节）、1ZB（Zettabyte，泽它字节）和 1YB（Yottabyte，尧它字节），每个单位之间的运算关系是乘以1024。截至 2009 年，美国几乎所有部门中每一个雇员数量在 1000 人以上的企业所存储的数据平均值至少为 200TB，是美国零售商沃尔玛 1999 年的数据仓库的两倍。很多经济部门中，每个企业平均存储数据超过 1PB。欧洲的组织2010 年存储容量总计接近 11EB，大约为整个美国存储容量（16EB 以上）的70%。全球企业 2010 年在硬盘上存储了超过 7EB 的新数据，消费者在 PC 和笔记本电脑等设备上存储了超过 6EB 新数据，而 1EB 数据就相当于美国国会图书馆中存储数据的 4000 多倍（James，2011）。数据容量增长的速度大大超过了硬件技术的发展速度，以至于引发了数据存储和处理的危机。大量的数据会被处理掉，比如医疗卫生提供商会处理掉他们所产生的 90% 的数据（手术

① 《IBM 大数据战略》，载《中国计算机报》，2011-11-07。

过程中产生的几乎所有实时视频图像）。

然而，大数据不只是大。海量数据引发的危机并不单纯是数据量的爆炸性增长，还牵涉到数据类型的改变，也即多样化（Variety）。原来的数据都可以用二维表结构存储在数据库中，如常用的 Excel 软件所处理的数据，我们称之为结构化数据。但是现在，更多互联网多媒体应用的出现，使诸如图片、声音和视频等非结构化数据占到了很大比重。有统计显示，全世界结构化数据增长率大概是 32%，而非结构化数据增长率则是 63%，预计至 2012 年，非结构化数据占有比例将达到互联网整个数据量的 75% 以上。用于产生智慧的大数据，往往是这些非结构化数据（曹磊等，2011）。

Informatica①中国区首席产品顾问但彬认为："大数据"包含了"海量数据"的含义，而且在内容上超越了海量数据，简而言之，"大数据"是海量数据 + 复杂类型的数据。但彬进一步指出：大数据包括交易和交互数据集在内的所有数据集，其规模或复杂程度超出了常用技术按照合理的成本和时限捕捉、管理及处理这些数据集的能力。

简单来说，大数据由三项主要技术趋势汇聚组成。一是海量交易数据：在从 ERP 应用程序到数据仓库应用程序的在线交易处理（OLTP）与分析系统中，传统的关系数据以及非结构化和半结构化信息仍在继续增长。随着企业将更多的数据和业务流程移向公共和私有云，这一局面变得更加复杂。二是海量交互数据：这一新生力量由源于 Facebook、Twitter、LinkedIn 及其他来源的社交媒体数据构成。它包括了呼叫详细记录（CDR）、设备和传感器信息、GPS 和地理定位映射数据、通过管理文件传输（Manage File Transfer）协议传送的海量图像文件、Web 文本和点击流数据、科学信息、电子邮件，等等。三是海量数据处理：大数据的涌现已经催生出了设计用于数据密集型处理的架构，例如具有开放源码、在商品硬件群中运行的 Apache Hadoop。Hadoop 是一种以可靠、高效、可伸缩的方式对大量数据进行分布式处理的软件框架。它的可靠性在于提前假设计算元素和存储会失败，因此它维护多个工作数据副本，确保能够针对失败的节点重新分布处理；高效性则表现在它以并行的方式工作，通过并行处理加快处理速度。Hadoop 还是可伸缩的，能够处理 PB 级数据。此外，由于 Hadoop 依赖于社区服务器，因此它的成本比较低，任何人都可以使用。对于企业来说，难题在于以具备成本效益的方式快速可靠地从 Hadoop 中存取数据。脸谱是 Hadoop 最知名的用户之一。通过 Hadoop，类似脸谱的社交网站

① Informatica，企业数据集成解决方案提供商，1993 年创立于美国加利福尼亚州，并于 1999 年 4 月在纳斯达克上市（纳斯达克代码：INFA）。已在国内成立北京、上海、广州、台湾及香港分公司。

和类似国内淘宝网的商业网站实现了“推荐你可能认识的人”、“可能想读的书”、“可能喜欢的商品”等服务。

1.3　大数据与云计算

我们再来了解一下与大数据密切相关的“云计算”概念。《互联网周刊》主编姜奇平认为，大数据并不像某些人说的，是云计算之“后”的又一浪，而就是云计算本身，因为两者都是数据的大规模集聚与定制化分布的结合。

云计算的基本原理是，使计算分布在大量的分布式计算机上，而非本地计算机或远程服务器中。云计算意味着只需要一台笔记本或者一个手机等互联网终端，就可以通过网络服务来实现我们需要的一切，包括浏览文档、图片、视频，甚至运行超级计算这样的任务。云计算的应用包含这样的一种思想：把力量联合起来，给其中的每一个成员使用。云计算的演进如图 1-4 所示：

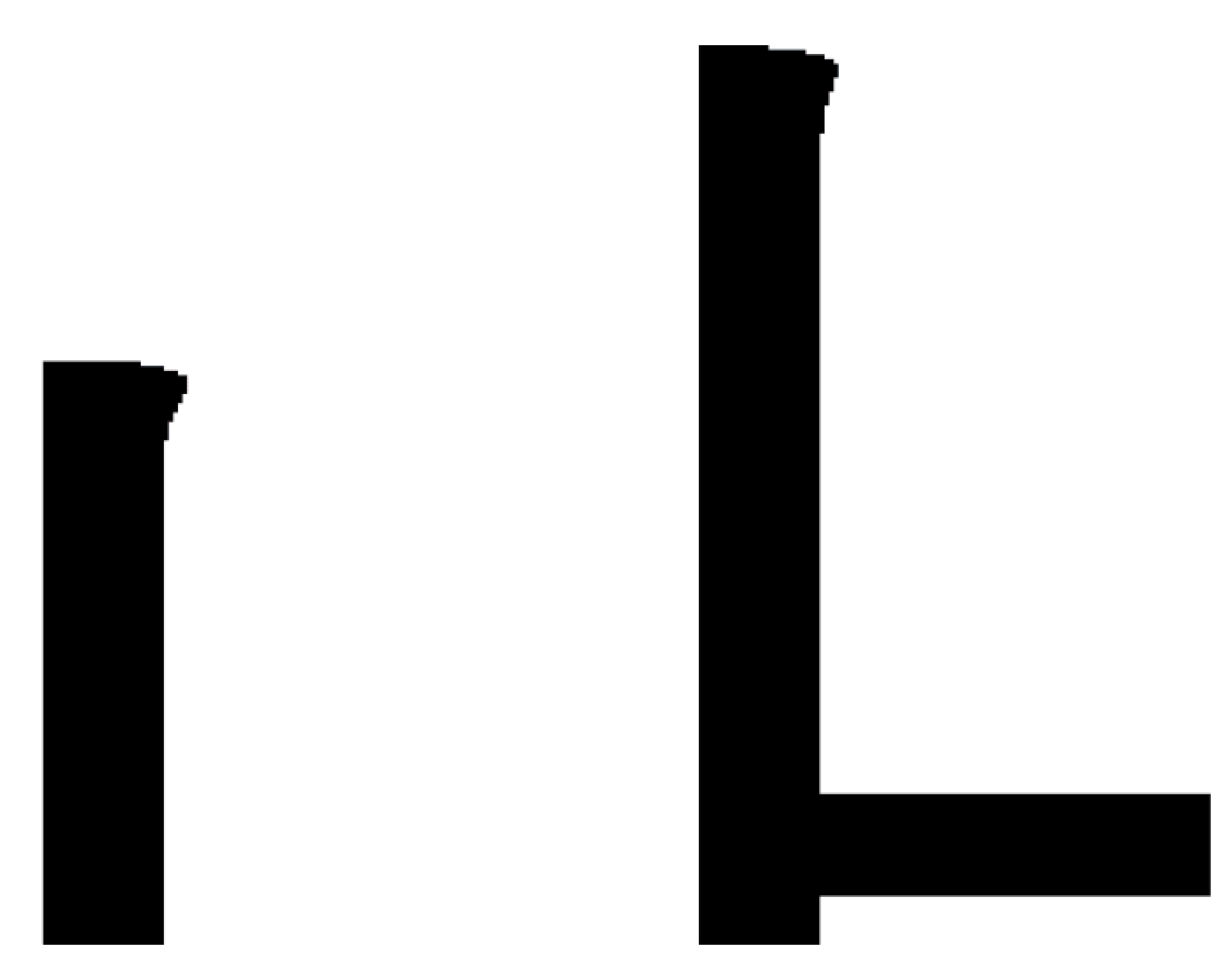


图1-4　云计算的演进①

终端用户通过网络浏览器或手机应用程序使用云计算，而这些商业软件和数据被存储于距离遥远的服务器上。支持者声称云计算使得企业的程序运行得更有效，更迅速。通过提高管理性能和缩减维护费用，并且对 IT 进行强化使其能更迅速地对资源做出调整，以满足流动性和不可预见的商业需求。

云计算的概念可以回溯到 20 世纪 60 年代，约翰·麦卡锡（John

① 图片来源：中国云计算网，http://www.cloudcomputing-china.cn/。

McCarthy）[①]曾预言“有朝一日，计算可能会变成公共设施”。几乎所有当今云计算的特征（弹性供给、作为公共设施提供、在线、无穷供给），道格拉斯·帕克希尔（Douglas Parkhill）[②]在他 1996 年出版的著作《效用计算的挑战》（*The Challenge of the Computer Utility*）中都已涉及。另外，他在书中还提到了云计算，电力工业与公共使用，私人、政府与社区使用形式的对比等内容。其他学者表示云计算的起源可以提前到 20 世纪 50 年代。当时，计算机科学家赫伯·格劳希（Herb Grosch）[③]假设整个世界将在由大约 15 个巨大的数据中心运作的终端上运作。

沃尔玛或者谷歌这样的大公司，很早就开始使用云服务，但是成本极其昂贵。今天的商业硬件、云建筑及开源软件正在将大数据带入普及民用的领域。即使是在车库中创业的公司也可以用较低的价格租用云服务了（Edd，2012）。云计算应用的快速增长得益于高效网络的普遍适应性、低成本的电脑、存储设备以及硬件虚拟化的广泛应用等因素。

如果说“云计算”是大数据时代必备的技术平台的话，那么“众包”（crowd sourcing）业务就是大数据时代全新的生产组织模式。这是《连线》（*Wired*）杂志 2006 年发明的专业术语，用来描述一种新的商业模式，即企业利用互联网来将工作分配出去、发现创意或解决技术问题。通过互联网控制，这些组织可以利用志愿员工大军的创意和能力——这些志愿员工具备完成任务的技能，愿意利用业余时间工作，满足于对其服务收取小额报酬，或者暂时并无报酬，仅仅满足于未来获得更多报酬的前景。尤其对于软件业和服务业，这提供了一种组织劳动力的全新方式。

“众包”模式使得科学发现不再是专业学者埋头于实验室的苦差事，而是全球科学家、学生和感兴趣的民众都可以参与的大众活动。谷歌公司在这个领域做了许多工作，他们开发了 Google.org——这是一个利用谷歌在信息技术处理数据方面特长建立的全球公众都能够参与的科学研究平台。从 2008 年 11 月起，Google.org 启动了名为“流感趋势”的项目，使用一种复杂的算法，对关于流感的网络搜索进行跟踪，从而对流感在人群当中传播的方式作出估计。其后，Google.org 还组织了地球引擎项目，将大量的卫星图像和数据开放给公众，让每个人都可以对气候影响下的水源变化和沙漠化进行跟踪研究。这些项目都在寻求一种“长尾效应”，用来解决过去一直无法展开研究的科学难题（曹磊

① John McCarthy（1927—2011），美国人，普林斯顿大学数学博士。计算机科学家、认知科学家，被誉为“人工智能之父”。

② 加拿大计算机专家，“云计算”概念的提出者。

③ Herb Grosch（1918—2010），提出“格劳希法则”（Grosch’s Law）。认为计算机越大越好，计算机的大小决定其运算速度。

等，2011）。

1.4 大数据的价值

麦肯锡全球研究所报告《大数据：创新、竞争和生产力的下一个前沿》（James et al., 2011）指出，对企业而言，大数据的价值体现在两个方面：分析使用和二次开发。大数据的分析使用可以揭示之前由于分析成本太高而忽略的信息，如消费者的同伴影响[①]、购物交易信息分析、社交网络信息和地理数据等。过去的 10 年中，大数据分析方面已经积累了一些开发新产品和新数据的成功经验，比如数据推送改变了 Facebook 一类社交网站的商业化模式，其 8 亿用户潜在的个人信息和商业价值都能被转换成各类广告用途，与广告主对接，哪怕用户的一句“最近胃疼”的状态更新，也会吸引胃药广告和保险广告的“轮番轰炸”。

对个人而言，智能手机的普及为开发大数据应用搭载了最好的平台。苹果公司 2011 年秋天发布了 iPhone 手机上的问答软件 Siri。这款软件源起于五角大楼的一项研究，之后却成了硅谷的一项创新产品。苹果公司 2010 年买入 Siri 技术，并不断加入更多数据。如今，使用苹果手机的人们提供了数以百万计的问题，Siri 正在成长为一个越来越熟练称职的个人助理，提供事件提醒、天气预报、就餐建议等服务，回答的问题也愈加广泛。

在种类繁多的大数据中，有一类是探测人们所在位置时产生的定位信息。GPS 技术的不断发展使得我们可以在几十米的距离内定位像手机那么小的装置，同时我们也看到了个人位置数据被用来创造新的商业和创新企业模式，这种模式几乎涉及全球每个人的生活。

个人定位数据涉及通信、零售、媒介等多个行业。这个领域蕴含着创造新价值的巨大潜力，麦肯锡全球研究所预测服务提供商将获得超过 1000 亿美元的收入，为消费者和终端用户创造的价值将达 7000 亿美元。

早期个人定位数据的来源是个人信用卡和借记卡付款，消费者在销售点终端（POS）的固定地点刷卡，与个人身份识别数据相关联。

随着手机用户增加，使用基站信号定位越来越普遍。现在，许多智能手机配备 WIFI 网络功能，这是用来收集定位数据的另一个来源。Skyhook 一类的服务商将不同 WIFI 网络的物理位置绘制在地图中，移动设备通过联入搜索到的 WIFI 网络，广播自己的位置信息。智能手机技术使个人位置数据更准确也

① 即 peer influence，指人们效仿同类中某种流行行为的影响力。

更容易获得，尤其对移动设备应用程序的开发人员来说。此外，正在开发的新技术甚至可以收集在 GPS 信号极弱的建筑内的个人定位数据。

如今，导航设备、基站跟踪、智能手机是大多数定位数据的来源。导航设备频繁更新位置数据导致数据量飙升；全球庞大的手机用户群产生的个人定位数据相当惊人；而智能手机用户使用的各种应用程序要求定位跟踪，也使得定位数据快速增长。来自麦肯锡的数据表明，2009 年全球范围生成的个人定位数据超过 1 拍字节，并且以每年约 20% 的速度增长。

日益增长的手机用户群每天都在生成庞大的数据，亚洲已成为产生个人定位数据的领先地区。2010 年，中国的手机用户世界第一，达 8 亿部；印度排第二，超过 6.5 亿部；北美大约是 3 亿部，位列第三。（见图 1-5）

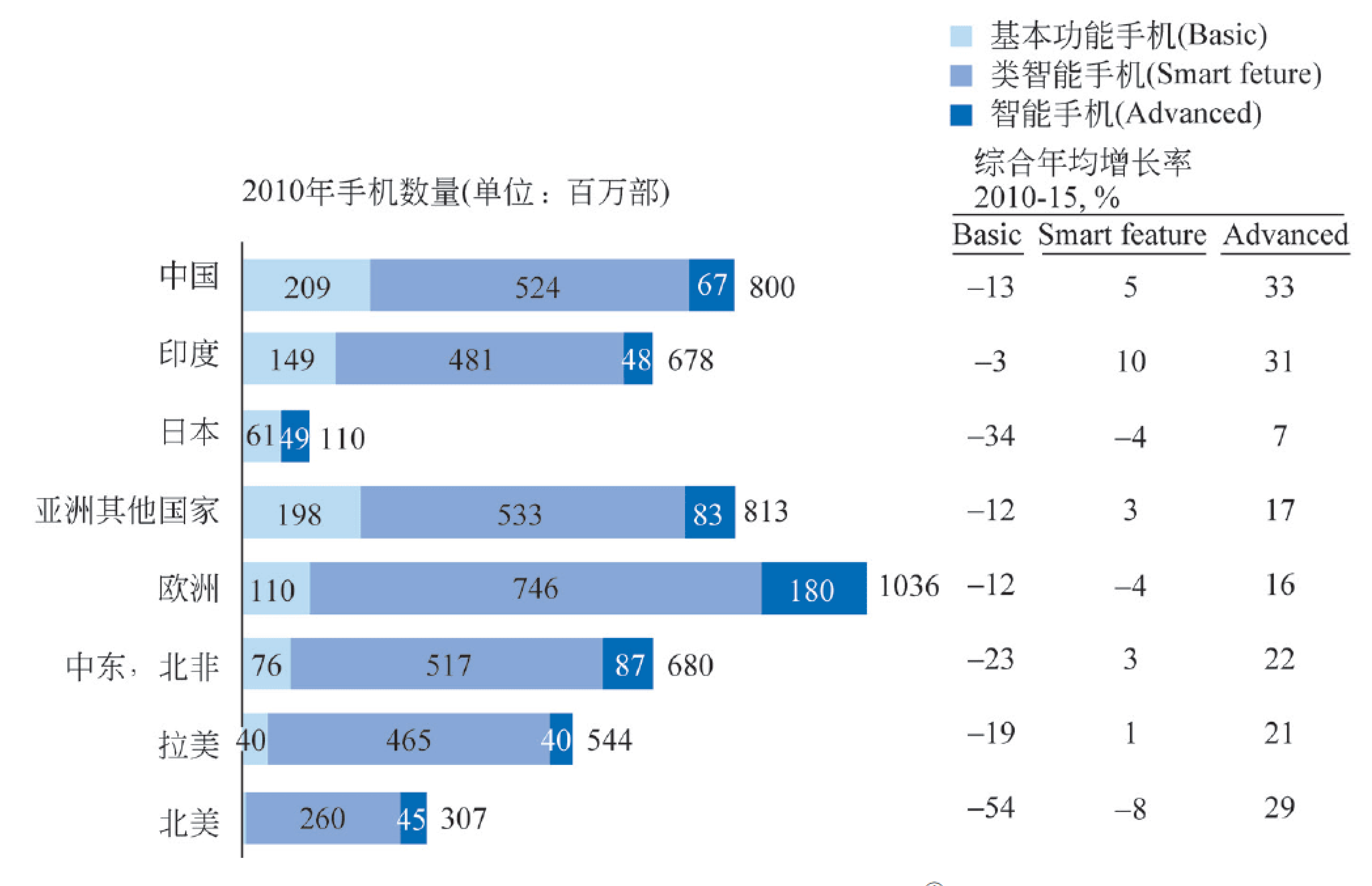


图1-5　移动电话在各国使用情况[①]

个人定位数据主要应用在以下三方面：一是个人使用的定位服务，包括智能路由、汽车车载智能通讯、智能手机移动定位等；二是个人定位数据组织性的使用，比如地理定位广告、电子收费、保险定价和应急响应；三是聚合定位数据在宏观层面上的使用，包括城市规划和零售业务智能。

定位数据体现了互联网技术与移动终端的不断融合；而“众包”这种大数据时代全新的生产组织模式，则影响着这个世界传统的思维逻辑。从 20 世

① 资料翻译自：麦肯锡全球研究所报告《大数据：创新、竞争和生产力的下一个前沿》。

纪 90 年代互联网搜索引擎投入民用以来，科学家就一直试图通过互联网在时时更新的线索中发现下一个全球流行疾病的线索。这个数据挖掘项目很长一段时间中都处在科学实验的阶段，直到 2008 年 11 月谷歌推出了流感趋势监测系统（Google Flu Trends①），这是一个运用类似数据“众包”处理方案（Crowd sourcing）预测流感暴发情况的在线数据处理平台。谷歌的科研团队发现，每周世界上都有上千万的互联网用户在线搜索健康信息。在流感高发期，以“流感”为关键词的在线检索会大大增加；过敏高发期，以“过敏”为关键词的在线检索增加；到了夏天，这一高频检索词则换成了“晒伤”……他们意识到也许是已经感染流感病毒的人群才会上网检索相关信息，在流行病疫情和在线检索行为之间或许存在某种关联性。尽管不是所有搜索流感信息的人都真的得了流感，但所有检索信息叠加就会呈现出一种趋势。谷歌的科研团队将通过分析网上检索信息得到的趋势与传统流感监测系统相比较，发现检索高频区正是流感暴发区。他们认为通过计算在线检索行为的频率，可以预测世界上不同国家和地区的流感暴发程度，并在 2009 年 2 月的科学杂志《自然》发表了这一题为《使用搜索引擎查询数据监测流感疫情》的研究成果（Jeremy et al., 2009）。2012 年，杜佳斯和她的团队将这一成果的数据与前去相关暴发地医院就医的数字进行比较，发现紧随网上增长的检索行为，短期内医院就医人数出现增加（Andrea et al., 2012）。

继“谷歌流感趋势”之后，还有美国的“健康地图”（Health Map②）和日本的“发现病毒”（Bio Caster③）等在线公共卫生数据挖掘平台投入使用。“健康地图”通过不间断地扫描博客、推特、官方监控数据、新闻网站和 RSS 链接及用户上传的信息，用十种语言发布监测结果，美国卫生和公共服务部借助与它的合作发布季节性流感趋势及 H1N1 病毒在美国的传播情况。约翰·布朗斯坦（John Brownstein）是哈佛大学医学院、波士顿儿童医院流行病学家，也是“健康地图”的联合创建人之一。他认为数据挖掘的意义在于将监测数据转化为有效的健康护理，“我们更需要看到（数据处理）产生的影响”。在他看来，学界对使用社交网站挖掘数据的信心在不断增加。2012 年 1 月，布朗斯坦和他的同事通过“健康地图”实现了对 2010 年海地地区霍乱疾病的动态监控，比当地的医疗工作者通过传统数据检测发布的报告提前了 2 周（Trop et al., 2012）。布朗斯坦表示，“随着我们使用数据挖掘技术发布论文的增多，相信会有越来越多的研究者会接受这种方法的合法性”。鲁米·川那拉（Rumi

---

① http://www.Google.org/flutrends/。

② http://www.healthmap.org/zh/。

③ http://biocaster.nii.ac.jp/。

Chunara）是哈佛医学院研究生物传感器的工程师。他认为非正式数据是传统监测方式的有益补充，尤其是在拥有较少健康资源但手机拥有率较高的地区和人群中。

“发现病毒”的开发者奈杰尔·科利尔（Nigel Collier）是位于东京的日本国立信息情报研究所的计算机专家。他认为对于自动化工作的软件来说，一大难点是如何从大量数字垃圾中解析出有用信息，他的主要工作就是通过软件过滤互联网上的垃圾信息。科利尔认为从自媒体采集的数据与官方发布的数据相比享有更多的优势，比如地理覆盖面更广、语义信息更丰富、成本更低。美国疾病预防控制中心、世界卫生组织、欧洲疾病预防控制中心和日本卫生部都和科利尔的“发现病毒”有合作，它们的成果包括追踪 H1N1 流行病和海地爆发的霍乱等。除此之外，科利尔的系统在监测互联网不发达的国家的其他疾病时也取得了很好的效果。和“健康地图”一样，“发现病毒”无偿发布信息，但其工作主要基于亚洲的新闻资源和搜索引擎，比如百度和搜搜。

“健康地图”和“发现病毒”这样的疾病跟踪服务系统，改变了世界卫生组织和美国疾病防控中心一类的卫生组织监控流行病暴发和应急反应的方式。林恩·菲内莉（Lyn Finelli）是亚特兰大疾病防控中心监控及暴发反应组组长。她说，“如果有人打电话说某个国家有学校因为流感暴发停课了，我们要做的第一件事就是在‘健康地图’上查询”，并表示，在疾病防控中心看来，数字疾病监测项目搜集到的信息可信度很高，也便于反馈。她同时认为新型数据监测的项目不能取代传统信息收集方式，但“有助于我们把力气使在对的地方”。

和社交媒体上迅速传播的话题相比，流行病的发生相对缓慢，即使是一种突然暴发的疾病，也需要几周才能确认病例。有些研究者开始使用谷歌新闻、推特或脸谱等社交网站实现快速追踪流行病，提供医疗援助。豪特（Taha Kass Hout）是位于亚特兰大的疾病防控中心信息科学部副主任，他认为“既然社交媒体已经存在，我们就要好好利用”。2012 年 2 月，《自然》发布消息称，这一技术目前已开始在全球范围内的公共健康机构中使用。2012 年 2 月，计算机界和流行病界的顶级专家，齐聚哈佛医学院参加数字疾病监测国际研讨会，讨论类似推特和社交媒体上的“非正式”信息如何快速地改变疾病监测（Rebecca, 2012）。

并非公共健康领域的每个人都准备好了迎接这场数据革命。安德烈·杜加斯（Andrea Dugas）是位于马里兰州巴尔的摩的约翰霍普金斯医院的急诊病学研究学者，他认为目前使用非正式数据工具需要审慎的态度，在这项技术成为决定公共健康相关政策的核心之前，还要进一步检测其可信度。哈佛大学公共卫生学院（波士顿）的流行病学家马克·利普斯琪（Marc Lipsitch）认为“数

据挖掘系统发现的是暴发情况，而不是具体案例”，数字跟踪平台只能提供补充信息，不能取代传统信息收集方式。

仅凭互联网上的日常闲聊，并不能解析出重要的疾病信息，这也是公共健康官员们认为网基协议（Web-based protocal）不能取代传统流行疾病评估的原因。目前鲜有证据能证明这些数据处理平台的稳定性和可信性。科学家们计划开展更大规模的研究，检测推特和谷歌“流感趋势”一类的数据处理平台能否在更广泛的地理区域内实现对多病种的预测。劳伦斯 • 马多夫（Lawrence Madoff）是位于伍斯特的马萨诸塞大学医学院传染病专家，他编写了“新出现疾病监控程序”（ProMED），这是世界最大的开源（Open sourcing）疾病暴发报告系统之一，但他指出是疾病防控中心通过传统的病例汇报发现了 2009 年的 H1N1 流感暴发，而不是在线监测系统。马多夫认为数据挖掘技术比以前有了长足进步，但仍然存在局限，他说“我们需要判断什么对我们来说是重要的”。

另外，即使流行病监控可以与现实同步，有效的健康服务仍然依仗财力和医疗资源。加拿大多伦多大学达拉拉娜公共卫生学院的流行病学家大卫 • 菲斯曼（David Fisman）认为，预测或者监控流行病的暴发并不能真正改变疾病防控的现状。

“谷歌地图”是谷歌公司提供的另一种“众包”业务。它为包括谷歌地图网站、谷歌搜寻器、谷歌交通等其他的地图程序提供支持，并且使地图成为三维形式，它提供世界各地的许多国家的街道地图路线，包括徒步、汽车、自行车或公共交通和城市商业定位。

“谷歌地图”提供了世界各地的城市地区高分辨率的卫星图像，各国的政府都在抱怨恐怖分子可能利用这些图像实施恐怖袭击。谷歌已经模糊了一些重要领域（主要是美国），包括美国海军天文台区（副总统的官邸所在地）和以前美国国会及白宫。其他知名的政府设施，包括 51 区在内华达州的沙漠都是可见的。谷歌的映射引擎促使民众对卫星图像的兴趣高涨。大多数高分辨率图像的城市是取自 800~1500 英尺的飞机空中摄影，其余大部分剩余的图像则来自卫星。虽然这些数据并不是随时更新,但有时这些数据也会和一些时间吻合,例如 2011 年 10 月 8 日，关于洛杉矶好莱坞的地图正好与奥斯卡典礼的摆设相吻合。

大数据的故事正在书写,却已带来了丰厚的经济回报。2004 年 8 月 18 日，谷歌上市，融资 17 亿美元，这个现在看起来不起眼的数字，在那个互联网泡沫破灭，投资者们谨小慎微，转而寻求医药、能源等领域投资机会的时代，已经称得上是一次了不起的成功。谷歌的上市重燃了全球不知多少 IT 青年的创

业梦——当年，扎克伯格就在哈佛大学的学生宿舍中创立了自己的公司。全球团购网站的鼻祖 Groupon 2011 年 11 月 3 日上市，融资额 7 亿美元，首日收盘报 26.11 美元，较 20 美元的发行价上涨 30.6%。中国社交网站人人网 2011 年 5 月 3 日上市，融资额 7.43 亿美元。2012 年 5 月 18 日，全球最大社交网络公司 Facebook 在纳斯达克上市，IPO 发行价为每股 38 美元，计划发行 4.21 亿股，估值达 1040 亿美元，成为美国有史以来上市时市值最大的企业。

1.5　大数据面临的大挑战

随着数据的增多，麻烦似乎也随之而来。迈克·斯通布雷克（Michael Stonebraker）[①] 教授今年 2 月在《美国计算机协会通讯》(*Communications of the ACM*）期刊发表了一篇题为《研究者的大数据危机》的专栏文章。文中描绘一位大学教授，调查了他周围 19 位同事，发现加上他自己，这 20 个研究团队所需要处理的数据都超过 100TB[②]，也就是说这 20 个研究团队的数据处理量几近 1PB[③]。普通的服务器根本承载不了如此庞大的数据集，而类似 EC2 一类的亚马逊弹性计算云服务（基于网络的计算服务，该服务可使企业用户在 Amazon.com 计算环境下运行应用程序）价格又过于昂贵，一般的科研项目也支付不起。斯通布雷克教授由此呼吁美国政府抽出“庞大运算项目”（Massive computing）中的小部分预算给“庞大数据管理”（Massive data management），从建立一个容量 400PB 的数据服务器开始，并交给真正懂得大数据的人管理。在专栏文章的最后他提出，“总得有个更好的法子”，因为“困难总是越来越突出”。

1.5.1　大数据意味着多学科集合

威瑞森通讯公司（Verizon Communications）[④] 的迈克·博迪（Michael L. Brodie）认为，当下我们面临的大数据挑战来自如何有效管理难以想象的海量数据以及如何将这些海量数据整合成我们所需要的有效信息，而不只是要酷似的玩弄技巧。鉴于计算机科学的每个领域都有着各自的理论和应用，多学科集

① Michael Stonebraker，数据库科学家，SQL Server/Sysbase 奠基人，于 1992 年提出对象关系数据库模型，在加州伯克利分校担任计算机教授达 25 年。现任 MIT 麻省理工学院客席教授。

② 1TB (Trillionbyte，万亿字节，又称太字节)=1024GB。TB 是现在电脑硬盘最大的存储量单位，10TB 大约相当于一个人脑的信息存储量。

③ 1PB（Petabyte，千万亿字节，又称拍字节）=1024TB。

④ 威瑞森通讯公司 1983 年成立于费城，是美国第一大地方电话公司和第二大电信服务商，拥有 1.12 亿固定电话接入线和 2800 万移动用户。该公司也是世界上最大的话簿出版和在线话簿检索公司。

合的“大数据”便有可能成为一种集成问题解决方案（Brodie et al., 2011）。例如，在美国的医疗保健大数据库（US Healthcare Big Data World）里，保存着散布在全美范围的5000万个病人的信息。“有效使用”的意思就是，如果输入“年龄在54周岁、高中辍学的白人女性，血压水平为150/80，β受体阻滞剂[①]治疗组，出现某两种并发症状，正在服用某三种药物”的信息，可以在数据库里匹配到另一个“年龄在54周岁、高中辍学的白人女性，血压水平为150/80，ACE抑制剂[②]治疗组，目前在服用同样药物”的信息（Begley, 2011）。

针对这一案例，“有效使用”可以分解为如下步骤：（1）明确问题。比如“针对54周岁女性高血压患者的有效药物”；（2）数据检索匹配。比如“所有的54周岁女性高血压患者”；（3）格式化（ETL）[③]信息，等待处理；（4）获取实际解决方案。实际解决方案可能涉及所有54周岁高血压女性患者的情况，进一步的挑战在于析出不同变量，比如成千上万个54周岁女性高血压患者的其他身体指标、社会网络、薪金、受教育程度等；（5）回答并解决问题，比如“获取有效的药物”。

事实上，在任何一个数据库中，非结构性数据（图片、声音和视频）所占比例都越来越高，数据储存量从早先的吉字节已经发展到了太字节、拍字节和艾字节，结构性数据在美国医疗保健大数据库中的比例已不足10%，并且这一比例还在急速下降。大多数关联子数据库的语义格式并不兼容，因此大部分的数据分析仍然需要人工，这是实现以大数据库为基础的“集成问题解决方案”的难点所在，也即博迪提到的语义（正确性）和工程（效率性）局限。

1.5.2 海量数据意味着增加了有效使用数据的难度

当下的Web 3.0时代是“基于数据的网络”时代（Web of Data），互联网已经成为一个超大的关系型数据库（表1-1）。其特征为：（1）个性为主；（2）强调用户体验；（3）良好的模块制定功能；（4）数据整合能力强（周珍妮，陈碧荣，2008）。据统计，现有数据网络含有310亿个RDF[④]三元组，其中4000多万个RDF链接的三元组将不同数据源之间的数据串接起来。这些数据中政府数据占41.9%、地理型数据占19.4%、出版和媒体类数据占14.8%及生命科学数据占9.7%。

① Beta blocker，β受体阻滞药，一种治疗高血压和心脏病的药物 。
② ACE inhibitor，血管紧张素转化酶抑制剂，一种高血压制剂。
③ ETL指Extract（析出）、Transform（格式转换）和Load（下载）。
④ Resource Description Framework，资源描述框架。

表1-1　Web 1.0、Web 2.0 与Web 3.0 比较（郑慧会等，2009）

Web 1.0	Web 2.0	Web 3.0
读，单向的、被动的接受信息	“写”和共同建设，仅基于服务商提供的平台	高度自主权，用户在互联网上的信息数据可以跨网站平台使用
机械化门户	半智能化 搜索 + 个人空间 + 门户	完全智能化 基于搜索 + 个人关键词标签 + 个人化空间 + 智能匹配的新门户
通过浏览器浏览网页	加上很多通过 Web 分享的其他内容，互动性更强	完全基于 Web，用浏览器即可实现复杂的系统程序才具备的功能

原始的大数据呈现出一片混乱的状态。从事数据工作的人普遍认为 80% 的精力都用在了数据清理上，正如彼特 • 沃登（Pete Warden）[①] 在其著作《大数据词典》中所言：“我可能花更多的时间整理那些杂乱的源数据，而不是直接就开始分析数据。”

数据网在以下三方面为数据整合和大数据处理增加了难度。一是通用和专有词汇的使用。像“人”、“产品”、“出版物”一类常见的表达，关联数据（linked data）[②] 资源可以借用；但其他常见表达里没有的词汇关联数据资源需要自定义。借用更多广泛运用的常见表达词汇，可提高不同数据资源的通用性。二是不同格式数据对同一对象描述的认定。不同计算机语言之间对同一对象的描述可能不同，比如 owl:equivalentClass，owl:equivalentProp-erty，rdfs:subClassOf，rdfs:subPropertyOf。应用程序如能辨识同一对象在不同语言中的表达，将有助于数据集合和数据清理。三是由于媒介平台的开放性，自媒体时代人人都在发布资讯（数据），大部分的互联网数据都是垃圾数据（SPAM），因此科学评估数据质量和确定有价值的数据子集也是一大挑战（Christian et al.，2011）。

1.5.3　语义网技术的广泛应用面临两大挑战

一是目前缺乏成功案例；二是从未消失的数据整合困境，使得不同数据

① 软件工程师，曾在苹果公司任职 5 年，创立网站 http://www.openheatmap.com/，实现电子数字表格可视化。出版两本大数据相关书籍：《大数据词典》（*Big Data Glossary*，*O'Reilly Media, Inc*.，2011）和《大数据手册》（*Big Data Handbook, O'Reilly Media, Inc*.，2011）。

② 万维网创始人 Tim Berners-Lee 提出，指语义万维网第一种可行的表达形式，实用且可操作，适用于各种形式的数据。关联数据是一组最佳实践的集合，它采用 RDF 数据模型，利用 URI（统一资源标识符）命名数据实体，发布和部署实例数据及类数据，从而可以通过 HTTP 协议揭示并获取这些数据，强调数据的相互关联、相互联系以及便于人机理解的语境信息。

之间的链接难以实现。数据整合是互联网行业最“烧”钱的领域之一，每年的投入超过上百亿美金。如果个人和企业可以从中获取价值实现盈利，那么技术的突破就不是问题了。LOD①数据抓取概念源于政府公开数据蕴含的无限商机（Christian et al.，2011），这一概念认为实现图片资源等非结构性数据的搜索是使用政府公开信息的有效途径。如果有一家网站可以实现编程语言的搜索、数据的搜索或是网站用户更早前上传图像信息的搜索，这将成为这一系统最主要的互动模式。要实现这样的功能，则需要图像处理、排序方法、视觉化、关键词匹配等各项技术的成熟运用。2011 年 6 月，搜索引擎网站谷歌（Google）推出了“以图找图”（Search by Image）功能。这一产品是利用图片内容、透视和颜色等因素进行图片搜索，以帮助用户找到近似的图片搜索结果。该技术采用了自动图片识别技术和元数据技术。Google 图片搜寻引擎除了可以让使用者添加图片网址来搜寻图片，也可支持图片上传，如果用户使用的是 Google 浏览器，还可直接用鼠标拖曳图片的方式，快速上传图片，直接搜寻图片。

1.5.4 大数据平台需要可以处理不同种类数据的数据整合技术

Openlink 公司② 的首席软件设计师奥瑞 • 俄凌（Orri Erling）指出目前人们已经意识到了智能数据处理的前景，但现实使用情况几乎还是空白。类似现在运用的 OWL 语言③ 可能是数据融合的处理方式之一，但不会是未来的方向。目前的关联数据和 RDF 在数据整合技术中占有一席之地，它们的国际通用性强，且为无预定数据模式。RDF 是 Resource Description Framework 的缩写，即资源描述框架，是一个用于表达关于万维网（World Wide Web）上的资源信息的语言。它专门用于表达关于 Web 资源的元数据，比如 Web 页面的标题、作者和修改时间，Web 文档的版权和许可信息，某个被共享资源的可用计划表等。然而，将“Web 资源”（Web resource）这一概念一般化后，RDF 可被用于表达关于任何可在 Web 上被标识的事物的信息，即使有时它们不能被直接

---

① 1976 年，Clark 提出了细节层次（Levels of Detail，LOD）模型的概念，认为当物体覆盖屏幕较小区域时，可以使用该物体描述较粗的模型，并给出了一个用于可见面判定算法的几何层次模型，以便对复杂场景进行快速绘制。

② OpenLink 公司是交叉资产交易、风险管理和操作处理软件提供商。

③ OWL（Web Ontology Language）是 W3C 开发的一种网络本体语言，用于对本体进行语义描述。由于 OWL 是针对各方面的需求在 DAML+OIL 的基础上改进而开发的，所以一方面要保持对 DAML+OIL/RDFS 的兼容性；另一方面又要保证更加强大的语义表达能力，同时还要保证描述逻辑（DL，Description Logic）的可判定推理。

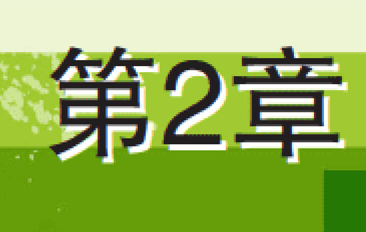
从 Web 上获取。比如关于一个在线购物机构的某项产品的信息（规格、价格和可用性信息），或者是关于一个 Web 用户在信息递送方面的偏好的描述。

RDF 使用 XML 语法和 RDF Schema（RDFS）来将元数据描述成为数据模型。数据对资源的描述是与领域和应用相关的，比如对一本书的描述和对一个 Web 站点的描述是不一样的，即对不同资源的描述需要采取不同的词汇表。一个 RDF 文件包含多个资源描述，而一个资源描述是由多个语句构成，一个语句是由资源、属性类型、属性值构成的三元组，表示资源具有的一个属性。资源描述中的语句可以对应于自然语言的语句，资源对应于自然语言中的主语，属性类型对应于谓语，属性值对应于宾语，在 RDF 术语中称其分别为主语、谓词、宾语。由于自然语言的语句可以是被动句，因此前面的简单对应仅仅是一个概念上的类比。RDF 规范并不定义描述资源所用的词汇表，而是定义了一些规则，这些规则是各领域和应用定义用于描述资源的词汇表时必须遵循的。通过 RDF，人们可以使用自己的词汇表描述任何资源，由于使用的是结构化的 XML 数据，搜索引擎可以理解元数据的精确含义，使得搜索变得更为智能和准确。如果 RDF 和标准化的 RDF 词汇表在 Web 上广泛使用，而且搜索引擎能够理解使用的词汇表，就可以避免当前搜索引擎经常返回无关数据的情况。

数据库管理系统（Database Management System），是一种操纵和管理数据库的大型软件，用于建立、使用和维护数据库，简称 DBMS。它对数据库进行统一的管理和控制，以保证数据库的安全性和完整性。用户通过 DBMS 访问数据库中的数据，数据库管理员也通过 DBMS 进行数据库的维护工作。DBMS 提供数据定义语言 DDL（Data Definition Language）与数据操作语言 DML（Data Manipulation Language），供用户定义数据库的模式结构与权限约束，实现对数据的追加、删除等操作。DBMS 允许多个应用程序和用户用不同的方法在同时或不同时刻去建立、修改和询问数据库，将应用于更广泛的领域。

21 世纪是大数据的世纪，而关于大数据的故事，才刚刚开始。

（本章编译者：刘娟，清华大学国际传播研究中心助理研究员，博士生）



第2章 大数据应用

大数据有多大？IDC在2006年估计全世界产生数据量为0.18ZB(1ZB=100万PB)，而截至2011年这个数字已经提升了一个数量级，达到1.8ZB，相当于全世界每个人一块100多GB的硬盘。这种增长还在加速，预计2015年将达到近8ZB（Gantz et al., 2011）。

大数据的一个明显特征是数据的社会化（Socialization of data）。从博客论坛到游戏社区再到微博，从互联网到移动互联网再到物联网，人类以及各类物理实体的实时联网已经而且还将继续产生难以估量的数据。

为了阐述大数据如何创造价值，本章通过五个领域：医疗与健康、数据新闻学、社会管理、经济管理、物联网，为读者展现一幅浩瀚的大数据景观。五个领域对大数据的使用在其复杂性和成熟程度方面有所不同，由此提供了不尽相同的实践经验。它们也代表了全球经济中多种多样的关键环节，包括全球流通的部门诸如制造业，以及非贸易部门如社会管理，以及产品和服务的组合。

2.1 医疗与健康

改革现有的医疗制度，削减医疗成本不断上升的增长率，同时还要维持现有的优势，这是全球各个国家社会和经济共同面临的关键问题。麦肯锡全球研究所的报告——《大数据：创新、竞争和生产力的下一个前沿》(*Big data: The next frontier for innovation, competition, and productivity*）详细地介绍了美国在医疗健康领域中信息化和大数据应用的发展趋势。

医疗是美国最大的经济部门之一，医疗行业占美国GDP的17%，雇用了11%的美国劳动者。近10年中，美国医疗开支年增长率为历史最高，接近5%（扣除物价因素），是高位国债的重要构成部分。日益老龄化的人口和更新、更

贵的治疗方法将会扩大这个趋势。目前，医疗系统在提高运行绩效和采用科技辅助过程方面落后于其他许多部门。问题的严重程度和推动变革的迫切程度亟须果断地制度策略，尽快开始提高生产率，并削减不断攀升的成本（James et al., 2011）。

鉴于此，使用大数据库作为工具，将会辅助产生更有效、更加经济的医疗政策，更好的产品和服务，提供新的商业模式。根据麦肯锡的预测，在医疗领域具备所需的 IT 和数据库投资、分析能力、隐私保护以及适当的经济激励机制的情况下，大数据的使用将在 10 年内让美国的医疗市场获得每年 3000 亿美元的新价值，其中 2/3 以全国医疗开支的削减形式出现。

医疗系统攀升的成本带来巨大的财政压力，这促使美国国内出现了前景广阔的试点工程，使用大数据和数据分析管理工具来获得中长期的价值。类似这样的创新项目中，就有美国的退伍军人事务部（VA）已经成功推出的数个医疗信息技术和远程病人监控项目。VA 的医疗系统普遍在如下几个方面胜过私营部门：遵照医生建议的病患照顾过程，坚持临床指导，实现更高比率的循症药物疗法。这些成绩大多要归功于 VA 以绩效为基础的责任框架，以及使用电子病历实现的疾病管理方法。

位于加州的综合管理医疗联盟凯泽集团早期就将临床数据和费用数据相结合，应用至关重要的数据库，发现了“万络”的副作用，最终使得这种药物退出市场。

不仅在美国，欧盟国家也在加大医疗数据的收集和使用。英国国家卫生与临床优化研究所[①] 率先使用大规模的临床数据研究新药以及现有昂贵的治疗方法的临床效果。该机构提供适宜的费用指导，还经常和制药及医药产业协商价格与市场准入的条件。意大利药物局收集和分析昂贵的新药的临床数据，这是国家的成本效益项目的一部分。卫生局能够为新药加上“有条件的报销”，然后根据它的临床数据研究结果重新评估价格和市场准入情况（James et al., 2011）。

根据麦肯锡的研究以及美国医疗市场的经验，医疗数据有四种主要来源，每一种都为某一部分人群所有。在此领域内，数据的碎片程度相当高。这四个来源分别是：临床数据、付款人活动（索赔）和成本数据、制药和医药产品的科研数据、病人的行为和情绪数据。一般来说，医疗服务提供机构拥有广泛

---

① 英国国家卫生与临床优化研究所（National Institute for Health and Clinical Excellence，NICE）是英国国家医疗服务系统（NHS）的组织，设在伦敦和曼彻斯特。NICE 成立于 1999 年 4 月 1 日，目标是确保每个英格兰和威尔士人平等享有 NHS 医疗的机会。NICE 制定指南，设定质量标准，管理国家数据库，为 NHS、当地权威部门和其他组织提供指南。

电子化的财务和行政数据，包括账单及患者基本信息。但数字化和聚合临床数据的电子化仍处于初期。预测高达 30% 的美国临床文本 / 数字数据——包括病历、账单、化验和手术报告的——还没有进行数字化。即使临床数据是数字形式，它们也通常为个人所有，没有得到共享。事实上，大部分临床数据都是视频和监控动态，是实时数据而没有储存。

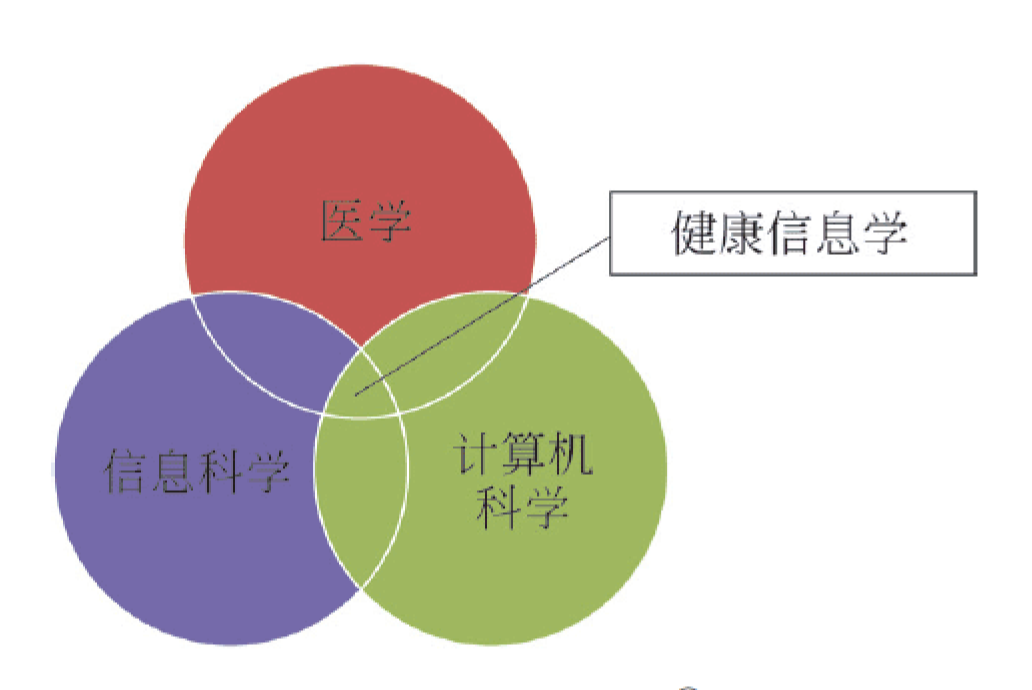


图2-1　健康信息学的构成[①]：医学、信息科学和计算机科学

美国及欧盟在临床、支付与定价、研究与开发、公共健康等领域中已经涌现出多种大数据技术，能够利用医疗部门中已有或可能获得的海量电子信息，提高医疗系统的效率和效果，比如费用削减、更高的效率、更好的治疗效果，以及生产力的提高。这些方法都需要对大数据库进行分析，这些数据库主要和医疗研发及供应相关，而不是关于医疗信息技术工具，比如处理医保申请的自动化操作（James et al., 2011）。

2.1.1　临床

在临床的范畴内有五种大数据工具，它们主要影响医疗服务供应者、支付者和医药制药公司提供临床治疗的方式。如果全部使用，这五种工具可以每年将美国医疗支出减少 1650 亿美元[②]（James et al., 2011）。

1. 疗效比较研究

结果导向的疗效比较研究（英文缩写为 CER），旨在通过分析详尽的患者和治疗结果信息，比较不同方案的效率，从而决定针对特定患者的最优治疗方案。许多研究显示，不同的医疗机构、地域和患者在治疗、结果和费用方面的差异非常大。分析包括患者特征、费用和治疗结果的大数据库能够帮助确定最有效和符合成本效益的疗法。如果医疗系统推行疗效比较研究，便有可能减少过度医疗和处理不足的发生率，这两者都会致使患者状况恶化以及产生更高昂的长期治疗费用。

在全球范围内，类似英国国家卫生与临床优化研究所的机构，如德国药物评估局（IQWIG），加拿大的统一药物评审（Common Drug Review），以及澳

① 资料翻译自：英国医疗信息化专家委员会，http://www.ukchip.org/?q=page/Professionalism-Health-Informatics。

② 基数是 2009 年的 2.5 万亿美元。

大利亚的药物福利计划，都开始成功实行 CER 项目。美国于 2009 年通过“美国复苏与再投资法案”，首次开始应用 CER。这项法案帮助建立了疗效比较研究联邦协调委员会，并获得 4 亿美元的拨款。为了发挥全系统的效能，还需要解决一些问题，比如收集和合并全面且一致性的临床数据集，使其为研究者可用。目前在施行 CER 的热潮中，仍然缺少标准和交互操作性，使得多个数据集难以合并。

2. 临床决策支持系统

第二项技术是使用临床决策支持系统提高手术及医嘱录入系统的效率和质量。目前此类系统可以分析医生的录入并将其和医学指导相比较，以便为可能的错误，比如药物不良反应或事故发出警报。通过使用医嘱录入系统，医疗服务提供机构能够减少不良反应，降低错误治疗和民事诉讼的比率，特别是降低医疗事故的发生率。一项在美国主要城市的儿科危症监护病房中的研究显示，临床决策支持系统在两个月时间内就将药物不良反应和医疗事故减少了 40%。

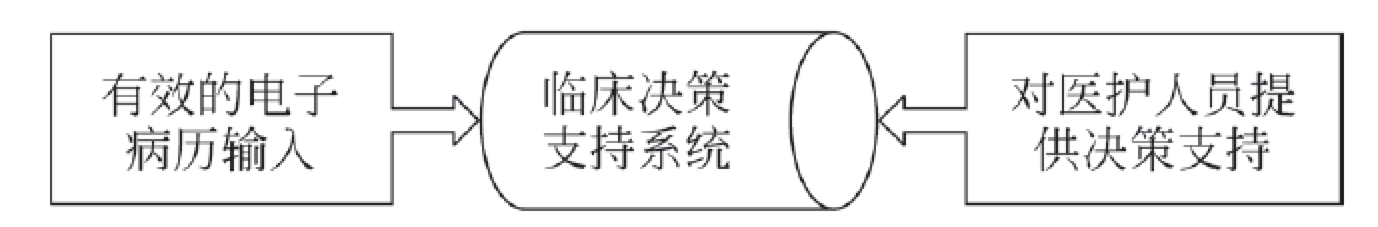


图2-2　电子病历系统（Electronic Medical Record，EMR）

未来类似这样的大数据系统将会变得更加智能，将会包含（X 光，CT，MRI）图像分析和识别的模块，还能自动挖掘医学文献以建立一个医学专业技术的数据库，并根据患者的病历向医生提出治疗方法的建议。此外，临床决策支持系统还能自动处理和帮助医生的咨询工作，让更多的工作量流向护理人员和医师助理，提升患者治疗的效率。

3. 医疗数据的透明度

临床大数据的第三种应用是分析关于医疗过程的数据，提高医疗数据透明度。这既能为医生和医疗机构指出提高工作水平的可能性，也能帮助病人挑选最合适的治疗方法。

医院信息系统（Hospital Information System，HIS），是增加医疗数据透明度的一项有效工具。在国际学术界，它已被公认为是新兴的医疗信息学的重要分支。HIS 系统的有效运行，将提高医院各项工作的效率和质量，促进医学科研和教学；减轻各类事务性工作的劳动强度，使他们节省出更多精力和时间服务病人；改善经营管理，堵塞漏洞，保证病人和医院的经济利益；为医院创造经济效益。

通过分析医疗机构的操作和绩效的数据集，可以创建进程图和仪表盘，让

数据透明成为可能。目标是确认和分析临床过程中差异与浪费的来源，让过程更加优化。记录医疗过程以及病人在医疗机构中的“路线”能够减少错误的发生。仅仅是公布费用、质量和绩效数据通常就可以形成竞争，促进绩效改善。这些分析可以带来机构改革，将会精简流程，节约成本，更有效地配备人员，提高医疗质量，改善患者体验，减少医疗费用。医疗保险和医疗补助服务中心正在测试“仪表盘”，这项创新将会实现透明政府的原则、促进公众参与以及合作。疾控中心也开始使用交互性的格式公布健康数据，提高性能以便处理数据。

公布质量和绩效数据还可以让患者了解医疗费用和质量的差异（目前大多仍是不透明的），做出更加明智的就医选择。数据的透明和适宜的报销计划将会鼓励患者在消费时更理性，这反过来促使医疗机构提高竞争力，最终提高整个行业的表现。

4. 远程患者监测

第四种临床大数据工具是通过远程监测系统收集慢性病患者的数据，分析结果以判断患者是否遵医嘱，以此改善用药和治疗方案。在 2010 年，美国约有 1.5 亿人患有慢性疾病，比如糖尿病、充血性心力衰竭和高血压，他们的治疗费用占到当年全国医疗费用的 80%（James et al., 2011）。远程患者监测系统对于治疗这些病人非常有效。该系统包括检测心脏的设备，可将血糖含量信息传递给看护者，甚至还包括“药片芯片”——当病人服用药片就发出报告的药物，几乎实时地将数据传给医疗记录数据库。一般来说，远程患者监测系统的数据可以减少患者住院时间，减少急诊，增进家庭陪护的匹配度，降低长期并发症。例如：向医生报告一位充血性心力衰竭患者因为水潴留而增加体重，便能够预防紧急住院。

5. 患者状况的高级分析工具

第五种工具是应用高级分析工具观察患者情况（比如分段和预测模型），确定那些能够从疾病预防和改变生活习惯中获益的人群。这些方法能够找出某种疾病的高风险人群，他们将会得益于预防性医疗计划；还能通过选择将已有症状的患者加入疾病管理项目，更好地满足他们的需求。患者数据还能提升衡量这些项目效果的能力。

2.1.2　支付与定价

这个类别中的两种工具都包含医疗支付和定价。支付和定价有潜力创造 5000 万美元的价值，其中半数来自于节省医疗开支的费用（James et al., 2011）。

1. 自动化系统

第一种工具是使用自动化系统（以神经网络为例的机器学习方法）识别欺诈，并核实支付者补贴申请的一致性和准确性。根据美国的支付者行业预计，每年补贴申请中的 2%~4% 是虚假或是不正当的；官方预计这笔费用高达医保和医疗救助的 10%。建立一个全面且一致的数据库，使用预算法来处理和检查申请的准确，检测可能性较高的诈骗、过失或错误，无论是实时的还是事后完成，都能够节省开支。如果实时操作，这些自动化系统能够在全额付款之前找出超额偿付，收回大笔损失。

2. 以卫生经济学和效果研究与绩效为基础的定价方案

第二种工具是基于真实的患者治疗效果数据，使用卫生经济学和效果研究与基于绩效的定价方案，实现公平的经济补偿——从支付给制药公司的药价到支付者付给医疗机构的偿付。

在药物定价方面，药厂将会共同承担治疗风险。对于支付者，一项重要的福利是新药的成本和风险分担计划，这能控制或限制相当大一部分医疗支付。同时，对医疗支付的限制也能使得制药公司获得更好的市场准入。它们还能够从更有效的用药方案（通过创新的定价系统而实现）中获得更高的利润。患者将能基于价格的公式集，以合理的价格购得创新药物，获得更好的治疗效果。为了让医疗系统实现最大价值，美国需要允许支付者的集体价格谈判。

以卫生经济学和效果研究为基础的药厂定价试点计划已开始施行，主要在欧洲。比如，诺华公司和德国健康保险公司达成一致，承担雷珠单抗注射液（Lucentis）每年超出 4.68 亿美元的支出，该药用于治疗与年龄相关的黄斑变性（James et al., 2011）。

2.1.3 研究与开发

在制药的子领域，五种大数据工具可以提高研发的生产力。它们可以共同创造高于 1000 亿美元的价值，其中 1/4 形式为更低的国家医疗费用（James et al., 2011）。

1. 预测模型

第一种工具是研究数据聚合，以便制药公司更好地为新药预测性建模，决定如何最有效率和符合成本效益地配置研发资源。“理性药品设计”意味着基于对临床前期或早期临床数据和研发价值链进行模拟与建模，从而尽可能迅速地预测临床效果。评价因素包括：产品安全性、疗效、可能的副作用、整体试验结果。这个预测模型可以在研发周期早期中止对次优混合物的研发和临床试

验，节约成本。

这种工具对于医药行业的益处包括：更低的研发成本，更精益、更快速、更有针对性的研发过程。它可以缩短药物问世的时间，创造出目标性更强的产品，扩大潜在市场，提高治疗成功率。预测模型能够将大约为 13 年的新药研发问世时间减少 3~5 年。

2. 统计工具和算法式改善临床试验设计

使用统计工具和算法式可以在研发过程中的临床阶段改善临床试验设计和招募患者的针对性。这个工具包括挖掘患者数据——评估患者招募的可行性、推荐更有效的设计、推荐有大量可选患者和优良记录的试验地点，以加快临床试验的过程。可以使用的技术有试验场景模拟，以及优化标签型号（适用于某种药物的适应症范围），这两者都可以增加试验的成功率。算法式将研发和试验数据与商业模型、历史监管数据相结合，找出针对试验的目标患者群体的规模和特征之间的最优平衡，以及监管部门对新药批准的可能性。分析还能改善选择研究员的过程——目标是那些经过证实有研究记录的人。

3. 分析临床试验数据

第三种工具是分析临床试验数据和病人档案，识别出药物的新用途并发现不良反应。在对大规模效果数据库进行统计分析、寻找出药物新用途的迹象之后，药物的重新定位或是新用途的营销成为可能。分析实时的不良反应病例报告让药物安全监视成为可能，使人们可以观察到常见临床试验中罕见的安全信号，识别出临床试验暗示出的但却没有足够统计解释力的事件。

这些分析项目在当下情境中格外重要，2008 年的年度药物召回创历史新高，而整体新药批准数量却在下降。药物召回通常对药企本身有很大伤害。2004 年撤销“万络”（一种抗炎症药物）的销售，导致默克公司耗费 700 万美元用于诉讼和索赔，其股东利益在短短几天之内下降了 33%。

4. 个性化药物

对新兴大数据库进行分析，是另一个很有前景、将能在研发领域创造新价值的大数据创新（比如基因组数据），将会提高生产力，研发出个性化药物。这个工具的目的是研究基因差异之间的关系、特殊疾病的易染病体质、特殊的药物反应，然后解释说明在药物研发过程中个体差异的原因。

个性化药物有希望在三个主要领域增进医疗水平：在患者出现病症之前进行早期检测和诊断；提供更有效的治疗，因为可以根据分子标记匹配细分有相同诊断结果的患者（即有同样疾病的患者通常对同样疗法有不同反应，这部分归因于基因差异）；根据患者的分子档案调整药物剂量，使副作用最小化，使

疗效最大化。

个性化药物正处于发展初期。尽管如此，它已经显现出惊人的初期成效，特别是胎儿基因测试的乳腺癌早期检测，以及白血病和结肠癌治疗的药剂测试。据预测，减少那些对个体患者没有疗效的药物处方可以节省 30%~70% 的开支。同样，鉴于肺癌早期手术费用大约是晚期手术的一半，早期检测和治疗也可以极大地减轻肺癌治疗对医疗系统的负担。

5. 分析疾病模式

与研发相关的大数据价值创造工具能够分析疾病模式和趋势，为未来的需求和成本建模，做出研发投资战略规划。这样的分析能够帮助药企最优化研发的侧重点，以此分配资源、设备和人力。

2.1.4 公共健康

大数据的应用能够改善公共健康监视和反馈。通过使用全国范围的患者和治疗数据库，负责公共健康的政府部门能够保证快速、协调地发现传染性疾病，全面监视疾病暴发，制订完整的疾病监测和反应计划。这项应用将会带来数不胜数的益处，包括减少医疗支出，降低感染事故，提高实验室能力，更好应对新发疾病与疾病暴发。

公共健康的研究者越来越多地采用地理信息系统（GIS）来分析人们所处的环境，以及这些环境如何影响个人健康。比如，GIS 中的道路网络数据可以提供关于某区域的交通拥堵情况、空气污染程度、城市化程度，并依据此分析该地居民的健康程度，比如心肺系统功能、心血管疾病和儿童肿瘤等（Frizzelle，2009）。有了准确和即时的公共健康报告，公众也会更加注意对和感染性疾病相关的健康风险，反过来降低传染的可能性。加在一起，这些因素可以创造更优质的生活。

2.2 数据新闻学

数据应用已经渗透到社会各个行业，这将对新闻学带来哪些影响？传统的新闻报道手法——采访目击者、讲述个人化的故事，在越来越趋向基于数据定量分析的时代，将会发生哪些变化？

在过去的几年里，全球一些具有创新精神的新闻媒体已经开始尝试利用数据更好地报道新闻，帮助读者理解正在发生的新闻事件，以及这些事件对人们生活的影响。这些尝试已经或多或少地改变了传统的新闻生产过程和呈现方式，

并将不可避免地对新闻学带来深远的影响。

欧洲新闻学中心（European Journalism Centre）和开放知识基金会（Open Knowledge Foundation）共同开发了一本《数据新闻学手册》（*The Data Journalism Handbook*），供全球用户免费下载和使用，旨在推动数据新闻学的发展。

《数据新闻学手册》是针对数据新闻这一新兴领域出版的一本免费开源的工具书。这本书最早开始于2011年设于伦敦的Mozilla Festival 48小时工作坊，尔后由来自澳大利亚广播集团、英国广播公司、《芝加哥论坛报》、德国之声、《卫报》、《金融时报》、《赫尔辛基日报》、《纽约时报》、美国在线新闻、《华盛顿邮报》、《芝加哥论坛报》、《世界之路报》、威尔士在线等诸多数据新闻领域的倡导者与资深专家以网络协作方式编写而成。该书目前仅有电子版，各个章节由不同的作者完成，网址为：http://datajournalismhandbook.org/1.0/en/。

数据新闻学是新闻学的新领域，这种新的新闻生产方式已经被英国广播公司（BBC）、《卫报》、《纽约时报》、《洛杉矶时报》等国际主流媒体广泛使用。欧洲新闻中心、Google等机构也从2011年开始，举办了各种有关数据新闻的全球性学术会议与竞赛活动。

陶氏基金会（Tow Foundation）与奈特基金会（John S. and James L. Knight Foundation）于2012年4月30日宣布将提供20亿美元来资助哥伦比亚大学新闻学院的数据新闻学研究项目，这项研究专注于数据新闻学的三个方面：

影响：衡量新的实践及工具如何影响受众及媒体资源；

新闻的透明度：关注公共数据——哪些是可用的，哪些不是；哪些是有用并与人们的生活息息相关的；

数据形象化：衡量哪种形式在传达信息与吸引读者方面最有效。

哥伦比亚大学陶氏数字新闻中心主管Emily Bell（2012）指出："大部分媒体仍不了解数据科学的发展前沿，以及信息传播对信息使用者的影响。我们旨在倡导那些对新闻学充满热情并具有相关知识的人才对数据新闻学开展研究。这不仅有利于解读大数据这一新领域，还能为新闻学在这个复杂多变的领域里提供指导。这项研究的目标是对新闻业和新闻学产生广泛的、直接的影响。"

在中国，数据新闻的发展方兴未艾。目前一些新闻传播院校已经开设相关课程，业界也有网易数读、政见CNPolitics[1]等，还有一些平面媒体、商业机构与个人，都在对数据新闻进行探索与尝试。

① http://cnpolitics.org。

2.2.1 什么是数据新闻学

数据新闻学（Data journalism）或称数据驱动的新闻学（Data Driven journalism），被认为是计算传播学的一个具体应用。通过挖掘和展示数据背后的关联与模式，运用丰富的、具有互动性的可视化手段，数据新闻学成为新闻学的新疆域和应用范例，并作为一门新的新闻分支进入主流媒体。比如，在猪流感疫情暴发时，每天都有从各地传来的最新数据，《卫报》网站的数据博客（Data Blog）设计了一幅猪流感疫情互动地图，展示世界各国的疫情进展。

所谓的数据新闻学，简单来说就是用数据报道新闻，它为记者将传统的新闻嗅觉与运用规模庞大的数据信息报道新闻创造了可能。

作为精确新闻学的进一步延伸，数据新闻学使新闻生产过程更为精细化，它对新闻工作者的技能要求除传统的文字写作、音视频制作外，还包括社科研究方法，计算机数据抓取、处理、可视化，平面 / 交互设计，计算机编程等多个领域。

数据新闻学是在多学科的技术手段下，应用丰富的、交互性的可视化效果展示新闻事实，把数据与社会、数据与个人之间的复杂关系用可视化手段向公众展示出来，以客观、易于理解的报道方式激发公众对公共议题的关注与参与。

任职于斯坦福大学的 Geoff McGhee 教授是一位以多媒体和信息图标为专长的记者，2009—2010 年，他在约翰・奈特新闻奖学金的支持下开始研究数据可视化。他认为，现在越来越多的新闻和数据有关，媒体的责任，是如何向公众解释复杂难懂的数据——既给予足够的信息，又不至于危言耸听。

McGhee（2012）制作了一则数据新闻学的教学视频——《数据时代的新闻学》，对数据新闻做出如下描述：

- 数据的爆炸式增长使得我们需要工具来进行分析。
- 可视化方面的专家正在开发工具帮助普通人更好地理解数据。
- 记者们则努力应对如何应用数据使新闻报道更加有说服力。
- 有经验的数据图表设计师能够把数据引入新闻学，但他们依然在论证数据对于概念诠释的有效性。
- 在一个连线世界中，数据越来越成为个人表达的载体。
- 数据将会实时推陈出新，极大地挑战着我们理解、分析和展示数据的能力。
- 创建在线可视化的技术正在转变，而新工具的出现将会使这个过程更加容易。
- 数据分析的重要性不亚于视觉展示，现有工具可以帮助实现这个过程。

2.2.2　数据新闻学的意义

当下，新闻故事不断涌现，它们来自众多信源：目击者以及博客，发生的事件在一个浩如烟海的社会关系网中被过滤、评级、评论，更多则是被遗忘。

因此，收集、过滤并将信息可视化的重要性日益凸显。人际网络、人—物网络之中使用的语言就是数据，那些在单个事例中无关紧要的微量信息，从全局角度看却有着非凡的重要性。现如今，一群具有开创精神的记者已经开始展示如何利用数据更好地理解我们当下所处世界中发生的事情，以及这些事件对我们生活的影响。

《数据新闻学手册》关于数据新闻的定义这样写道，数据新闻与其他类型的新闻区别在何处？或许在于将传统的新闻敏感和使用数字信息讲述一则好故事的能力相结合而带来新的可能性。这些可能性会出现在新闻报道的任何一个阶段：使用电脑程序自动处理信息收集和组合的过程，这些信息来自政府、公安局和其他公民机构（Bradshaw, 2011）。

数据新闻能够帮助记者使用数据图表讲述一个错综复杂的故事。比如，Hans Rosling①使用 Gapminder 软件将世界贫困进行可视化处理，吸引了来自全世界的关注；David McCandless② 广受欢迎的大量数据提取——比如政府开支的背景资料，或是冰岛火山造成和阻拦的污染物——显示出明确而清晰的设计的重要性。

数据新闻还可以帮助解释新闻事件和个人之间的关联，比如 BBC 和《金融时报》定期制作关于财政预算的互动性报道（观众可以发现预算对于自身的影响）。它还可以推动新闻采访的过程本身，如同《卫报》成功使用数据博客分享数据、背景和议题。

数据可以成为数据新闻的信源，或是讲述新闻故事所使用的工具，也可以两者兼得。和任何其他信源一样，数据应该接受怀疑和质疑；我们应该意识到它如何塑造并限制利用数据生产出的新闻报道，正和任何其他工具一样。

《数据新闻学手册》的作者们认为，通过数据的使用，记者工作的重点从“第一个报道者”转化成为对特定事件的影响的阐释者。话题的范围宽且远：下一

---

① 汉斯·罗斯林（Hans Rosling, 1948 年 7 月 27 日）是卡罗琳学院的国际卫生学教授，并担任 Gapminder 基金会董事长，该基金会开发了 Trendalyzer 软件，具有把统计数据图形化的特点，方便人们理解数据资料。

② 大卫·麦克坎德雷斯（David McCandless）曾为《英国卫报》、《连线》、《独立报》等刊物撰稿，擅长以简洁精美的图像展现复杂、抽象或分散的资讯，并将不同的数据组合，展现其中的联系和模式。大卫认为，数据可视化不仅是在信息丛林中找到方向的最好方法，还能帮助人们发现全新的视角。他的新作《信息是美丽的》（*Information is beautiful*）以其擅长的可视化数据描绘了当今世界的各个方面。

次正在酝酿的金融危机，我们使用产品背后的经济学，资金的滥用和政治决策失误，一些抽象的社会问题，比如失业如何影响公众——基于他们的年龄、性别、教育水平。使用数据能够将抽象概念转化为普通人可以理解并且会涉及的事物。记者们还可以分析复杂局面如骚乱和政治辩论中的动态关系，显示其中的谬论，帮助人们寻找复杂问题的解决方案（Lorenz, 2011）。

此外，深入的数据新闻提供更深刻的观点。如今，编辑室数量削减，大多数记者希望改行进入公关行业。数据记者或是数据科学家已经成为相当抢手的员工——不只在媒体行业。全世界的公司和机构都在寻找“意义制造者”，即那些掌握挖掘数据、将其转换为有形信息的专家。

数据大有希望，这激发了编辑室的兴趣，促使他们寻找新型的记者。对于自由撰稿人来说，熟练的数据操作同样提供了通往新职位和稳定收入的道路。不妨这样思考：与其雇用记者用低价值内容迅速填充版面，不如使用数据创造出对交互式内容的需求，而实现后者的唯一渠道是花费一整个星期研究一个问题。这在媒体行业中的许多领域都广受欢迎。

2.2.3 数据新闻学的功能

斯坦福大学 Geoff 教授（2009）曾长期担任《纽约时报》等媒体记者，他于 2009—2010 年间开始研究数据新闻。他指出：现在的新闻越来越多的和数据有关，媒体的责任是如何向公众解释复杂难懂的数据。数据的爆炸式增长使我们需要工具来进行分析，数据可视化专家正在开发工具帮助我们更好地理解和使用数据，记者的工作是运用数据使新闻报道更加有说服力。

1. 讲故事的新工具和新方法

数据新闻最重要的一项功能是使用数据可视化软件，通过统计大量的数据，帮助记者使用数据图表讲述错综复杂的故事，而这种讲故事的方式必须依赖于对大数据的分析和可视化处理。由于数据量巨大，按照传统的新闻生产方式是几乎不可能实现的。数据新闻最佳的阅读载体是交互性强的电子媒介（例如：接入互联网的电脑、手机终端等）而非传统的平面媒介。

例如，Gapminder 基金会创始人汉斯·罗斯林（Hans Rosling）教授等开发的数据可视化软件 Gapminder 在公共卫生、环境、公共安全等报道领域的应用十分广泛。以艾滋病报道为例，Gapminder 基金会使用该软件对 2007—2009 年世界各国人均收入与艾滋病（HIV）感染率进行了统计分析，如图 2-3 所示：

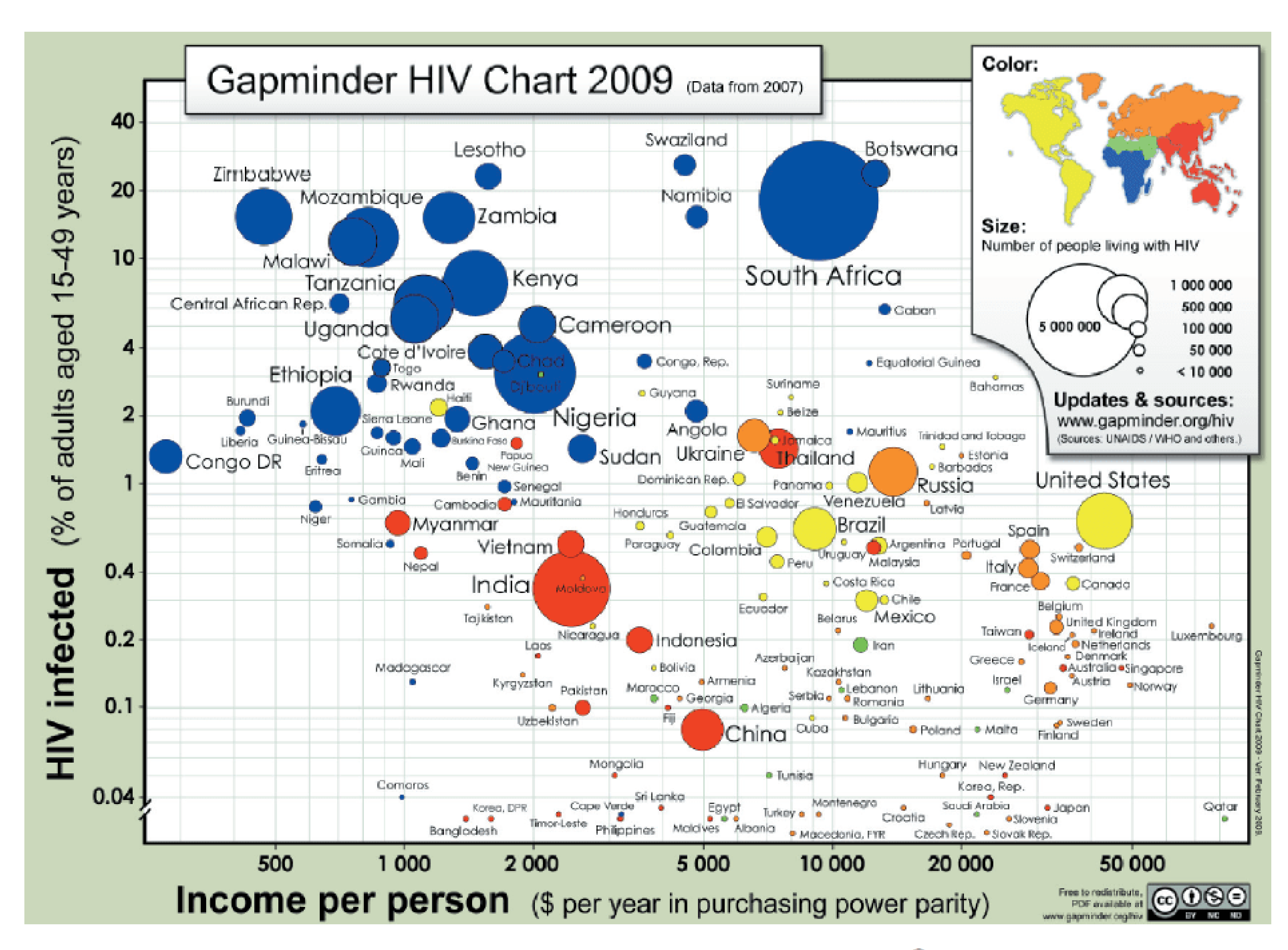


图2-3　Gapminder对艾滋病的报道[①]

在图 2-3 中，横坐标表示同等购买力下的人均收入（单位：美元 / 年），纵坐标表示艾滋病（HIV）感染率（15~49 岁成人感染百分比），图表主体中用气泡代表某个国家，气泡的面积大小表示艾滋病感染者存活人数的多少。

通过这幅图可以让读者一目了然地认识到艾滋病感染率与国民人均收入之间的线性关系：从大体趋势上来看，随着人均收入的增加，艾滋病的感染率也随之降低。读者如果通过电脑、手机等终端阅读这条新闻，还可通过交互功能了解更多的信息。大量庞杂的数据、变量之间的复杂关系，很难依靠传统的报道手法来讲述。大数据的应用与呈现在这则案例中已经不仅仅是信源的角色，更多的是承担讲故事的工具和方法。

2. 解释宏大新闻事件与个人的关联

数据新闻还可以帮助记者解释宏大背景下的新闻事件和个人之间的关联，新闻学对于报道公共事务的要求，是能够通过记者的报道，帮助读者认识到一项公共政策的实施或修订对个人造成的影响。数据新闻让读者们在阅读报道后对自己的生活提出问题，诸如：我们的家人是否安全？我们的孩子接受的教育是否合适？总之，数据新闻的工作是让读者能在数据和新闻事件中找到属于自己的故事。

① 资料来源：http://www.gapminder.org/downloads/gapminder-hiv-chart-2009/。

例如：英国广播公司（BBC）和毕马威会计师事务所联合制作的《预算计算器：2012 年的财政预算将如何影响你？》（*Budget calculator: How will the Budget 2012 affect you?*）能够帮助读者理解新的财政预算（税收计划）对个人生活带来的影响，用户只需要在界面上输入一些个人信息（例如：每周购买多少啤酒、多少包香烟、家里有几辆汽车、月收入多少等），它就能够自动计算出你需要为新的政府财政预算增加支付多少税，你的生活会变得更好还是更糟。如图 2-4 所示。

Budget calculator: How will the Budget 2012 affect you?
How much better or worse off will you be in the coming tax year following the Budget?
Use our Budget calculator, developed by accounting firm KPMG LLP, to find out how the measures being brought in next month will affect you.
1 Alcohol & cigarettes
2 About you
3 Income
4 Your car
5 Results
Alcohol & cigarettes
Estimate how much beer, wine, spirits and cigarettes you would consume in a normal week. You may choose to answer for yourself or your household.
You will be about £0.00 better off in 2012/13
Pints of beer 0
Glasses of wine (175ml glass) 0
Glasses of spirits (50ml doubles) 0
Packs of cigarettes 0
NEXT >

图2-4 BBC的财政预算计算器①

3. 记者角色的转变

通过数据的使用，记者工作的重点从“第一个报道者”转化成为对新闻事件的影响的阐释者。数据新闻报道的议题范围十分宽广，记者更有意义的工作是为读者提供经过定量分析的洞见，使用数据能够将抽象概念转化为普通人容易理解的事物，帮助记者讲述抽象的社会问题。记者们还可以分析复杂形势中各种变量的动态关系，能够为读者预见下一次正在酝酿的金融危机，指出政府对资金的滥用和决策的失误，甚至帮助人们寻找复杂问题的解决方案。

① 资料来源：http://www.bbc.co.uk/news/business-17442946。

数据新闻需要计算机程序员、数据分析人员与编辑、记者密切配合。《芝加哥论坛报》的新闻应用程序编辑 Brian Boyer（2011）说：“所有关于新闻应用程序的创意都来自新闻编辑部里的记者和编辑，我们之间建立起了非常牢固的个人以及专业关系，他们（编辑和记者）获得数据后，会向我们提供想法。”

程序员的工作主要是辅助记者，帮助他们挖掘数据，将庞杂的数据转化为电子数据表等。程序员实时处理新闻编辑部正在进行的数据工作，将其转化为应用程序——一张地图、一个图表，或是一个网站。

数据新闻的应用程序一般会占据新闻页面最显著的位置，和记者的文字报道相辅相成。

例如，英国《卫报》（*Guardian*）在 2011 年的伦敦骚乱中运用数据新闻的方法，帮助读者更好地理解了事态进展和背后原因。

伦敦骚乱发生后，英国政治上的保守派指责脸谱（Facebook）、推特（Twitter）和黑莓信使（BBM）等社交媒体传播谣言、煽动骚乱，并据此要求暂时关闭社交媒体，但政府没有调查骚乱发生的真正原因。

《卫报》与学界进行合作，邀请曼彻斯特大学的学术团队一起研究社交媒体在骚乱中的作用。后者一共分析了 260 万条关于骚乱的推特信息，观察谣言如何在推特上传播，不同的用户在宣传和散布信息中的功能，以确定推特和其他社交媒体是否煽动了骚乱。

《卫报》的“解读骚乱”数据团队使用地图显示骚乱发生地点的贫困程度（如图 2-5 所示），让“骚乱与贫困没有关系”的主流政治话语不攻自破。他们还制作了一段视频，将暴乱发生地和参与群众的家庭住址联系起来，显示出“暴乱通勤路线”，建模预测暴乱者最有可能采取的路线。此外，研究者对推特信息进行了内容分析，分类编码为：重复、驳斥、质疑、评论，并对数据进行可视化处理，指出了推特在纠正谣言方面发挥了积极作用。

2.2.4　数据新闻的采集和发布

1. 数据收集

《数据新闻学手册》为我们提供了一些简单的搜索建议。搜索数据时，需要包含数据的内容以及信息的格式或是来源。谷歌和其他搜索引擎提供文件类型分类搜索。现在的网络技术允许我们进行精准搜索，比如电子数据表（在搜索时附加“格式：XLS 格式：CSV”），地理数据（“格式：SHP”），数据库抽取（“格式：MDB，文件类型：SQL，文件类型：DB”），或是 PDF 文件（“格式：PDF”）。

另一种方式是从网络的专用数据端口、数据中心以及其他数据站点获得数据。

Mapping the riots with poverty

Data journalist Matt Stiles has taken our data on deprivation - and the riot incidents over the last few days and mashed the two up together. The darker reds represent poorer places, the blues are the richer areas. What do you think? Is there a correlation between the two?

• Interactive map of the riot events

• More on how we mapped deprivation

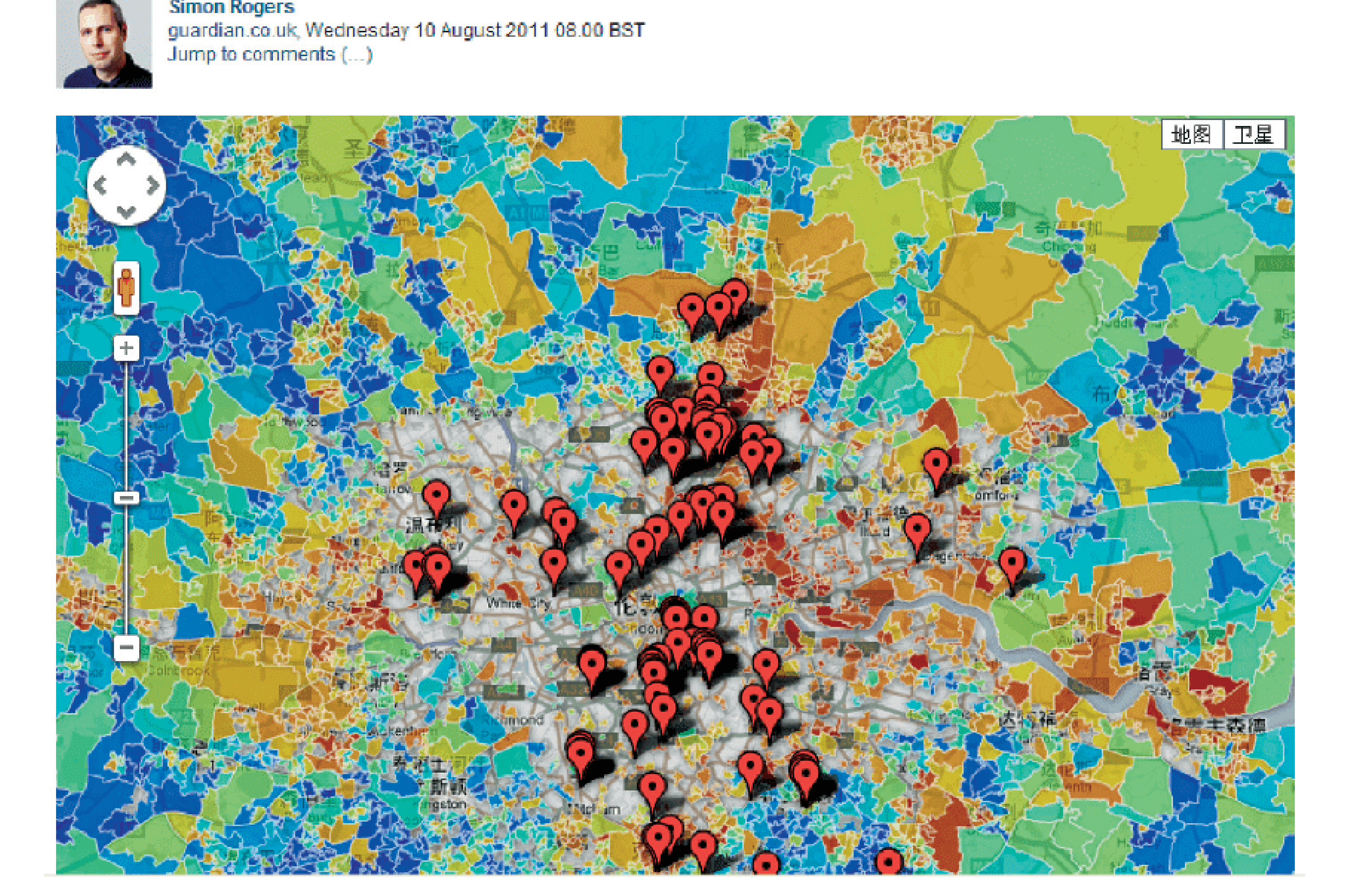


图2-5 《卫报》制作的地图显示骚乱与贫困的关系[①]

- 官方数据：越来越多的国家开始建立自己的数据门户，以促进公众和商业机构对政府信息的重新利用，诸如美国政府的 data.gov 和英国政府的 data.gov.uk。datacatalogs.org 提供了此类数据的全球最新索引。英国《卫报》的全球政府数据是一个元搜索引擎，包含许多国家的政府数据分类目录。
- data.hub 是由开放知识基金会运行的数据资源站点，便于查询、分享并重新使用那些已公布的数据。
- scraperwiki 是一个网络工具，可以更方便地提取碎片化的数据并在其他程序中重新利用，或从记者和研究人员那里检索数据。绝大多数数据是公开的，可以再次使用的。
- 世界银行和联合国的数据端口提供了所有国家的高水平指标，通常是多年的数据。
- 出现了一些致力于数据销售和再销售的创业公司，包括 Buzzdata——私人和公共数据包分享及合作和数据商店（如 Inforchimps 和 DataMarket）。
- Datacouch：这是一个上传、提炼、分享和可视化处理数据的地方。

① 资料来源：http://www.guardian.co.uk/news/datablog/interactive/2011/aug/10/poverty-riots-mapped。

- Freebase 是谷歌的一家子公司，由开放数据的爱好者们组成，提供人、地点和事物的实体图。
- 研究数据：国家和学科的数据集合不胜枚举，比如英国数据档案。

网络论坛是搜索数据的另一去处。GetTheData 是一个问答网站，包含各种关于数据的问题，包括：如何寻找关于特定题目，如何从一个指定数据来源获取信息，可视化工具，数据清理或转变为可操作的格式。数据记者还可以使用 Data Driven Journalism List 和 NICAR-L 的邮件列表，列表中汇集了众多数据记者和精通计算机辅助报道（CAR）的“极客”的信息。

除此之外，程序设计员们还组成了一个数量急剧增长的国际草根新闻组织，拥有几十个分部，成员数以千计，来自五湖四海。他们的共同目的是建立涵盖数据记者和技术专家的工作网络，重新思考新闻和信息的未来发展方向。

教授、政府公务员和业界人士都是可以提供帮助的群体。

2. 数据呈现

从原始数据配上新闻故事，到创造美丽的可视化和交互式网络应用程序，向公众展示数据有很多不同的方式。

有的时候，数据讲述新闻的效果胜过文字或照片，这就是“新闻应用”和“数据可视化”在编辑室里如此受人瞩目的原因。新闻工具和技术的大丰收（通常是免费的）也激发了人们的兴趣，它们可以让最不善于技术的记者们将数据转化为视觉故事。

像谷歌融合表、Many Eyes、Tableau、Dipity 和其他工具，让用户创建地图、图表、图形等数据图形变得易如反掌，此前只有专家才可以完成这些任务。记者所面对的问题则是是否应该将数据可视化。不适宜的数据可视化在许多方面效果适得其反。

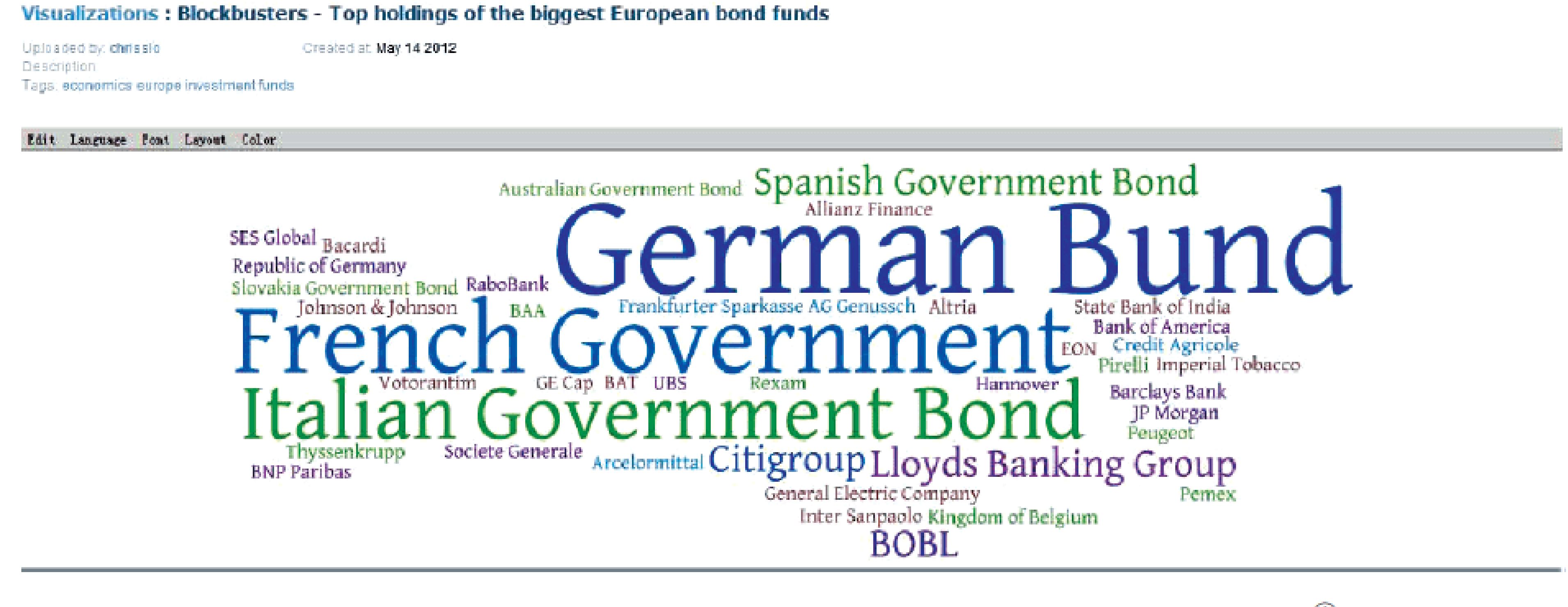


图2-6　应用Many Eyes制作的欧洲公债基金主要持有者[①]

① 资料来源：http://www-958.ibm.com/software/data/cognos/manyeyes/。图中主体部分文字为欧洲公债基金持有者机构名称，未翻译。

《西雅图时报》的 Cheryl Phillips 介绍说，数据记者们经常通过可视化处理将数据嵌入新闻，让读者可以轻松下载数据集，在可视化程序中进行互动或是利用数据本身挖掘新闻背后的更多事实。这样的新闻向读者公开数据，供批评者和更多感兴趣的人使用，显示出数据记者和编辑工作的透明度，他们也可以从批评者和读者那里获得更多建议。这些对于提高新闻质量十分重要。

《纽约时报》研发团队的数据设计师 Jer Thorp 的观点是，关于大数据的很多讨论遗漏了一项：人性面。人们大多把数据视为分离的、自由流动的数字，而忽略它们其实是对（通常是很有人性的）真实事物的测量。数据和真实的人、真实的生活紧密相连，数据专家必须要思考生产数据的真实世界。目前为止，位置数据的使用者都是第三方——程序开发员，知名品牌和广告公司。“第二方”（电信商和设备管理者）拥有这些数据，而“第一方”，即我们每个人既无法得到数据也无法支配这些信息。《纽约时报》的研发团队推出了一个叫 Openpaths.cc 的原型设计，允许公众探索他们自己的位置数据，并亲身体验数据拥有者的概念。毕竟，人们应该对和他们自身生活及经历密切相关的数据有一定控制权。

新闻应用程序允许透过新闻故事观察背后数据的通道。它们既可能是可搜索的数据库，还可以是漂亮的可视化产品。无论采用怎样的形式，新闻应用程序的主旨是鼓励读者和数据互动，发掘数据的意义，比如查询所处地区的犯罪趋势、社区医生的安全执业记录，或是他们自己选出的候选人的政绩。

2.3 社会管理

2.3.1 社会管理的运行面临重大考验

根据贝克的风险社会理论，不同于过去面临的大自然所构成的严重威胁，高度发达的工业社会带来的巨大风险和灾难不仅对整个部门和地区的财产、资本、就业机会、工会的力量等构成严重威胁，还对整个部门和地区的经济基础、民族国家的社会结构、全球市场等构成严重威胁。

这意味着从工业社会向风险社会过渡时将会经历一段制度混乱的过程，使得整个社会缺乏稳定感，而在这种制度混乱和社会动荡的状况下，所有的决定与决策都与地方政府有关，从行车速度限制、停车场建设、工业产品的生产细节等一系列具体政策，到政府有关能源供应、法律法规、技术进步等根本问题的总政策，都会突然被卷进与风险和灾难有关的根本冲突之中。

因此，风险社会对社会管理提出了严峻的考验。全世界范围内的政府都在努力应对日益沉重的压力，提高社会管理工作的效能。特别是在后经济衰退的余波中，许多政府都必须保持高水准的社会管理能力，寻求财政预算紧缩以减少国债，斥巨资刺激国内增长。除了降低债务水平，许多国家还面临中到长期的预算紧张——主要原因是人口老龄化将会大大增加医疗和社会保障领域的支出。

当前我国社会处于经济快速发展期，同时也凸显各种矛盾。政府需要收集巨量数据与数百万公民打交道，绩效表现往往参差不齐。尽管潜在好处巨大，但政府面临利用这一宝库的巨大障碍：很少有管理者主动发掘所拥有的信息，而政府往往将信息保存在各自为政的部门中。

政府部门是否可以通过大数据的应用提升自己的生产力和工作效能呢？麦肯锡研究了欧盟国家的政府部门行政管理，发现大数据的应用工具可以为社会管理提供有效的策略和技巧，以提升生产力、提高效率及影响力：欧盟政府部门可能会减少 15%~20% 的行政开支，创造 1500 亿到 3000 亿欧元的新价值。大数据还可以在未来 10 年中将年度增长率最高提高 0.5%（James et al., 2011）。

2.3.2　应用大数据推动社会管理

麦肯锡全球研究所的大数据研究报告显示，欧盟国家对大数据工具的应用可以从五个主要方面推动社会管理水平（James et al., 2011）。

1. 实现信息透明

若政府部门大数据库的数据更加易得，外部利益相关者（比如公民和企业）和内部利益相关者（比如政府雇员和政府机构）都能够提高自身的工作效率。比如，政府机构通过多种管理文件常规性地收集关于个人和企业的大量数据，但个人和企业则经常需要填写已经收集过且已经储存的数据。如果政府机构可以在收集数据时提供预先填写好的表格（已在政府数据库中的信息将会自动登记），表格提交者节省时间，政府机构也无须重复输入数据。目前越来越多的不同层级的政府部门开始引入“开放数据”的原则，允许公众获得原始政府数据。例如：美国的 data.gov、英国的 data.gov.uk 和西班牙的 www.proyecctoaporta.es 都属此类。个人和企业可以使用越来越大的数据库。这样的努力开启了海量的数据创新，人们将多种来源的数据结合起来（比如来自执法部门和市政工程的“官方”数据及来自社交媒体的公民记者“非官方”报道相结合）以创造类似“网络城市”新闻，记录在某个特定城市发生的事件。其他的案例包括

expectmore.gov 和 Dr. Foster，为英国公民提供卫生医疗信息，旨在直接测量项目的绩效。

2. 发现需求、展现差异和提高绩效

大数据的重要贡献之一是它可以发现不同政府机构在行使相似职能时呈现出的巨大绩效差异，这个信息对在机构内提高各部门的执行能力提供了重要机遇。比如，绩效仪表板显示出运行和财务数据，让政府机构衡量和比较各个部门的绩效。有研究者认为缺少外部竞争是政府行政水平较低的原因，这自然是原因之一。然而，即便在外部竞争压力较弱的情况下，凸显不同部门工作绩效差异仍可以带来内部竞争，提高效率。即使没有财务上的奖励机制，位于平均水平之下的部门负责人也会因为位列榜尾而希望有更好的表现。

3. 人口细分和定制政策

在私营部门，使用市场细分为个体提供定制服务的做法已经延续多年。然而，社会观念却认为政府部门应该为全体公民提供均等化服务。麦肯锡的研究报告发现，根据个体和人群将公共服务进行细分与定制能够提高效率、效果和公民的满意度。比如，德国联邦劳工局分析了数量巨大的历史数据，包括失业工人的历史、政府干预及其结果、求职花费时间等。随后劳工局根据此分析形成了人群细分，调整了政府对失业人群的帮助。这个政策连同其他措施实施了三年，帮助劳工局每年减少 149 万美元的开支，减少了失业人口重新入职的时间，而且提高了使用服务者的满意度。同样，政府的税收部门可以使用大数据对个人和企业纳税人进行分割。比如，可以将他们按照地理、守信记录、违约风险、征缴难度以及收入水平和人口特征将纳税人分类。有效的分割可以将潜在征缴和实际征缴之间的差距缩小 10%，同时更加精准的互动还可以将用户满意度提升 15% 之多。

4. 使用自动计算代替或辅助人为决策

大数据的更为复杂、更为高级的应用是使用自动算法来分析大数据库，从而帮助决策者判断。举例来说，政府机构需要找出财政支出中的异常，比如劳动部或社保部需要了解缴税、保险支付的异常状况。税收机构使用自动运算对纳税申报单进行系统和多层级检查，并且能自动标识出需要进一步检查或是审计的税单。这种方法能够大大促进征税工作的效能。

运算法则能够从多种源头抓取大量数据，识别出不一致、错误和虚假信息。比如，基于规则的算法能够标志出可疑的相关事件——一个人在收到失业补助的同时还提交了一份工伤案件。使用更加先进和调优的运算法技术（比如人工神经网络）可以降低错误判断和错误否定的可能性。在使用自动分析方法

之后，德国劳动部汇报减少了 20% 的错误发放补助。

5. 大数据应用在社会管理中的发展潜力

在欧洲国家政府部门的运行中，大数据的应用可以带来三个方面的回报：运行效率提高减少开支，减少出错成本和福利管理中的诈骗，以及缩小税收缺口。麦肯锡预测，提高效率的大数据应用适用于 20%~25% 的运行预算，可节省 15%~20% 的开支；减少福利发放中的错误以及通过欺骗获得的福利大约可以节省 40% 的成本；至于增加税收，预计税收缺口占欧洲税捐收入的 5%~10%，其中的 20% 可以被回收。总的来看，欧洲最大的 23 个国家政府可以在未来时间中创造 1500 亿欧元至 3000 亿欧元的新价值。

2.3.3　大数据对中国社会管理的意义

在矛盾多发期的中国当代社会中，公共管理和公共服务的维护事关全体社会成员根本利益，它的实质是保障和改善民生，维护社会稳定。目前的社会管理存在的制约要求管理部门形成科学有效的利益协调机制、诉求表达机制、矛盾调处机制、权益保障机制和统筹协调机制。

中央对于社会管理的高度重视体现在：2011 年 2 月 19 日，胡锦涛总书记发表的讲话中提及，社会管理事关科学发展，事关国家长治久安。2011 年 2 月 20 日，周永康评价社会管理“事关党的执政地位”。同年 5 月 30 日的政治局会议上指出，我国既处于发展的重要战略机遇期，又处于社会矛盾凸显期。社会管理理念思路、体制机制、法律政策、方法手段等方面还存在很多不适应。

运用社会大数据进行舆情研判，是社会管理创新的重要手段。

舆论安全是我国非传统安全的重要组成部分，舆论是可量化、可统计、可识别、可引导的意见流，舆论工作的重点是识别社会风险、掌握社情民意、支持科学决策，其政治性、专业性、技术性、系统性强。

随着互联网信息传播的全面普及，社会舆情越来越多地借助互联网进行传播，社会舆情从酝酿到爆发需要的时间越来越短，如果借助传统的社会调查等手段采集和分析舆情，往往贻误战机。因此，在新的传播形势下开展社会舆情研判和预警工作的着力点在于：以网络信息文本挖掘和分析技术为手段，通过建立灵敏高效的工作网络和科学完善的工作机制，总结舆论传播规律，及时识别风险并发出危机预警，并提出行之有效的危机应对策略建议。

截至 2011 年年底，我国已有网络舆情相关文献近 900 篇，主要集中在信息科学、社会学、公共管理学、新闻传播学等学科领域。

其中，信息科学的研究主要集中于如下内容：文本挖掘在网络舆情分析中的应用研究，互联网络舆情预警机制研究，互联网内容及舆情深度分析模式研究，基于主题聚类的热点发现研究，基于情感计算的网络中文信息分析及技术研究，基于信号分析的舆情预警研究，语义倾向分析方法研究，网络传播的无标度特征及其衰减规律研究，网络舆情监测系统中的主题帖自动标引及情感倾向分析研究等。

社会学和公共管理学的研究主要侧重于网络舆情传播对群体性事件的影响模型和动力机制研究，突发性事件的舆论管理，突发性事件的群体心理和行为，群体性突发性事件的网络舆情演变机制研究等。

来自新闻传播学的研究主要从舆论传播规律、舆论研判的指标体系、舆论预警指标体系等方面开展研究。

总的来讲，信息科学侧重于互联网文本挖掘和分析技术层面的研究，社会学和管理科学侧重于突发群体性事件管理中的群体心理行为和舆论控制研究，新闻传播学侧重于对舆论的本体进行规律性的探索和研究。

清华大学国际传播研究中心李希光教授主持的国家社科基金重大项目成果——社会舆情研判预警系统（以下简称舆情系统）是以社会舆情的监测、研判和预警为工作目标，以互联网信息挖掘技术和分析技术为基础，以计算机软件为主要工具，以灵敏高效的工作网络机制为保障，为党和政府的舆情管理等相关部门提供服务的工作机制。其目的在于及时了解和把握社情民意，对当前社会热点话题进行科学分析和研判，尽量降低各类突发事件带来的负面影响，科学预测重大危机事件的舆论走势，提供危机管理和应对的决策参考。

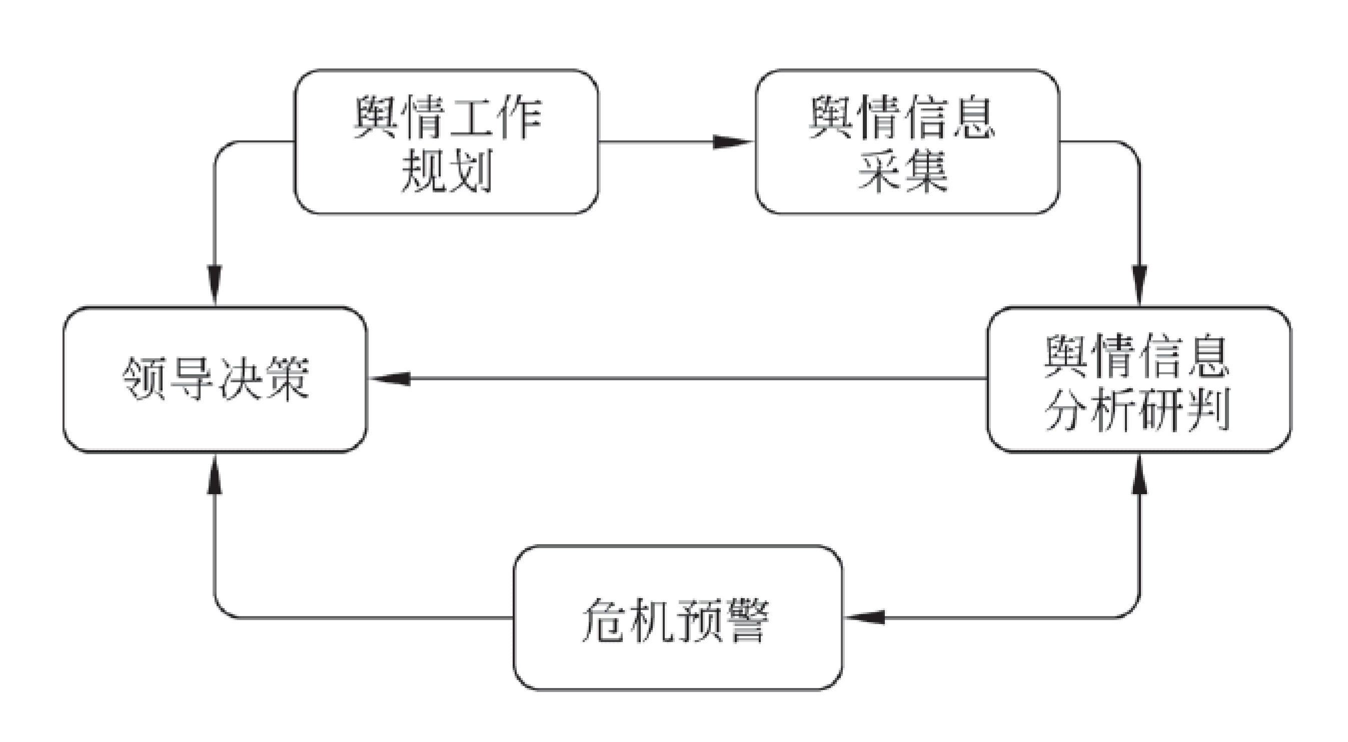


图2-7　舆情信息工作流程示意图

在舆情系统的工作流程中，领导决策既是舆情信息工作的起点，也是舆情信息工作的中枢，一切都是为了给决策者提供科学的舆情信息情报。为了使决策者充分和有效地使用舆情信息，有关舆情信息工作人员需要较高的政治觉悟，精准的分析问题能力，高度的危机意识，准确领悟舆情监测的内容，科学设计

舆情工作的规划，包括：通过哪些渠道调研舆情，调研哪方面的舆情，使用哪些关键词进行监测，监测的周期，监测的内容等。

舆情信息的采集如果通过人工的方式进行，既耗费大量的人工，还需要很长的调研周期，在信息瞬息万变的今天，这种传统的舆情采集方式越来越捉襟见肘，因此，本课题的主要产出——舆情系统，可根据预设的关键词和监测范围，通过接入互联网，自动采集新闻网站、论坛、博客、微博等多媒体平台的舆情信息，自动筛选有用的数据，自动统计，自动识别，大大地节省人力物力，为实时掌握舆情动态、发出预警提供了基本的条件。

网络舆情的研究属于跨学科重大研究课题，因此在研究过程中充分融合了新闻传播学、社会学、信息科学、管理学等学科的研究方法，从新闻传播学的实证研究范式出发，关注舆论酝酿、发酵、爆发的动力模式，以及在这一系列过程中信息的变化规律，以信息技术的最新进展作为手段，充分利用计算机和互联网技术中的文本挖掘技术、互联网爬虫技术[①]、自动聚类技术、自动标引技术、情感判断技术等，实现信息采集、信息存储、信息预处理、信息统计与分析的自动化。在此需要强调的是，上述技术的目的是为了实现对海量信息的快速、准确处理，以弥补人工处理信息的天然不足，但其基本思想仍然是来源于传统的新闻传播学研究方法，如：内容分析法、语义分析法等。

舆情系统的架构设计、工作流程与相应的关键技术解决方案如表 2-1 所示。

表2-1　舆情系统架构、流程与关键技术解决方案

舆情系统架构	工作流程	关键技术解决方案
信息采集系统	信息定向采集	基于网络爬虫技术的互联网信息采集技术
元搜索系统	信息全网采集	元搜索引擎技术（Meta-search Engine）
数据处理系统	信息预处理	对信息进行结构化预处理，文本挖掘技术
智能分析系统	定量描述与风险识别	舆情走势模拟，舆情热点发现，舆情态度分析，重点人物关联分析，重点机构关联分析
风险预警系统	发出预警信号	宏观舆情风险指数研究，微观敏感舆情识别研究

项目实现的基本技术路线是，使用面向对象的技术进行系统设计和实现，使用 Java 技术，遵照 J2EE 标准，其体系结构自下而上分为三层，分别为数据层、业务逻辑层和表现层。

① 网络爬虫，又被称为网页蜘蛛，网络机器人，是按照一定的规则自动抓取互联网信息的程序或脚本。

系统架构的示意图如图 2-8 所示：

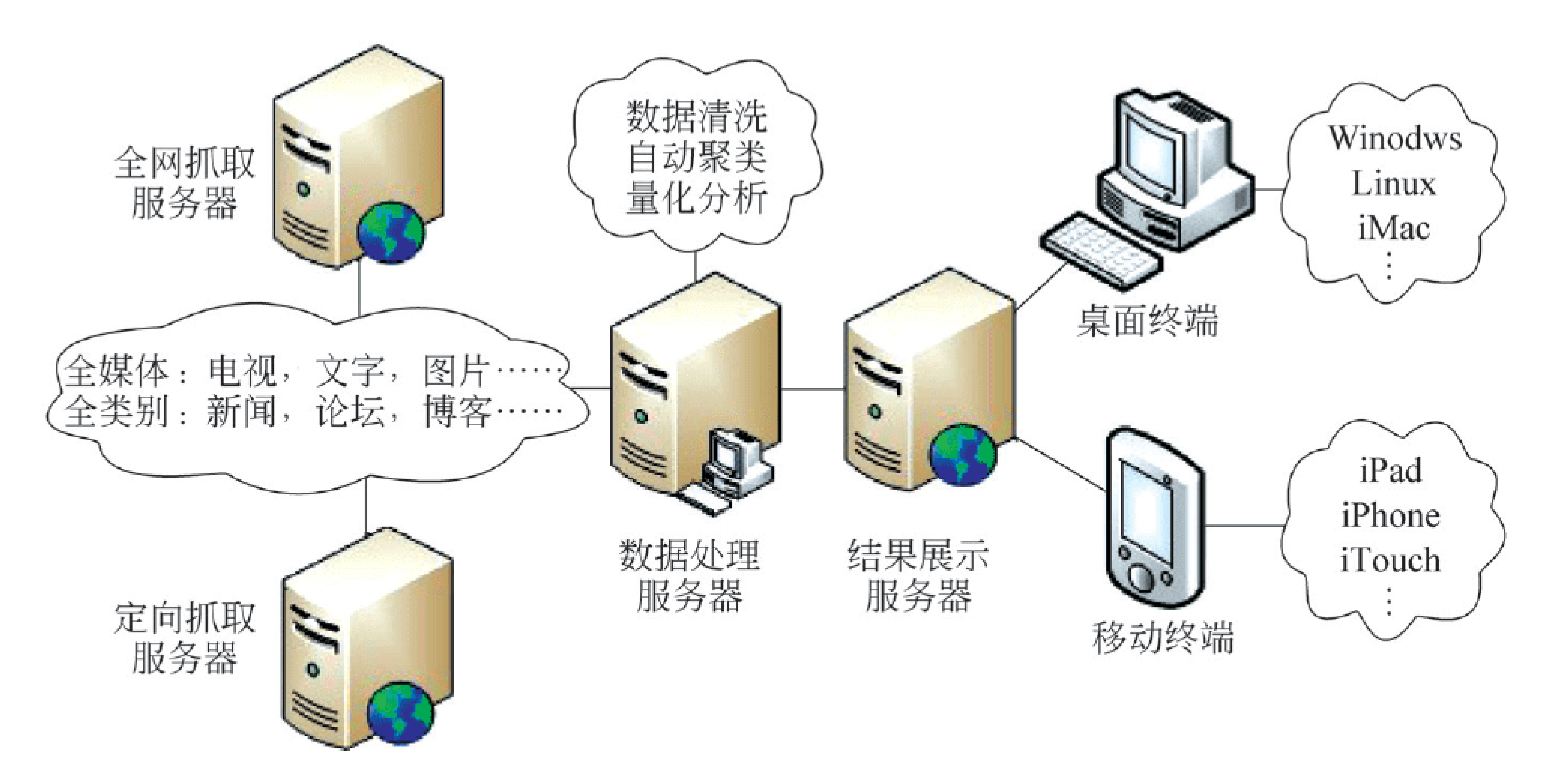


图2-8　系统架构示意图

此架构符合目前主流的舆情系统架构标准，不但能够很好地满足对系统的业务需求，而且具有较好的扩展性和安全性，具备强大的技术升级能力。系统的抓取服务器、数据存储服务器、运算服务器、展现服务器可分布式处理，能满足对大型数据业务的需求，可以在各种主流硬件平台和 Windows、Linux、Mac 等多个操作系统上运行，支持各类 Web 服务器和浏览器，通过各类电脑操作系统或各种终端，都可以访问、管理整个系统平台。

2.4　经济管理

2.4.1　零售业

和大数据相关的技术为创造价值提供了重大的新机遇。零售部门永无止境地争先发现和使用大数据为自己赢得竞争优势。零售商们不仅记录下每一笔交易和操作，还记录着新出现的数据源比如 RFID 芯片，可追踪货物、在线消费者的行为和情感表现，这使得数据量的增长势不可当。麦肯锡全球研究所的大数据报告介绍了大数据技术在零售业、制造业中的应用及其对整条行业链的影响（James et al., 2011）。

事实上，零售业通过使用信息技术的影响力获利的做法已经有几十年的历史。比如，在美国，零售终端的交易数据——主要从条形码中获得——在 20 世纪 70 年代首次出现。20 世纪 90 年代之后，许多大型零售商都开始使用门市层级和供应链的数据来优化配送和物流，加快货物规划和管理，升级店铺运

营。比如沃尔玛施行不间断的管理创新方法，直接和间接地促使了整个日用百货行业在 20 世纪 90 年代的生产力加速提升，比如仓储式格式，每日最低价，提升竞争强度，鼓励最优的管理和技术的扩散。沃尔玛还开拓了电子数据交换系统，将供应链用电子化的方式连接。沃尔玛研发的“Retail Link”可以让供应商大致浏览其门店，了解什么货品需要重新进货而不是被动等待订单。这种“厂商管理存货”的方法是一个革新性的概念，在 20 世纪 80 年代开始采用。这两种创新方式大大地提高了零售商的资本和人力的生产效率。当其他零售商在 20 世纪 90 年代开始模仿沃尔玛的首创，以保持竞争力时，整个行业的生产效率随之全面提升。

今天，领跑者们正在挖掘消费者数据，为从管理供应链到推销和定价等一系列问题提供决策参考。沃尔玛详细的、符合成本效率的消费者跟踪系统让零售商们可以挖掘消费者偏好和消费行为的大数据，从而从消费者产品生产商那里赢得关键的定价和配送特许权。

整个行业的零售商对大数据的处理变得日臻娴熟，数据来自多种销售渠道、商品目录、商店、在线互动。消费者数据日益颗粒状，而这些数据的广泛应用让零售商能够提高市场营销的有效率。将大数据工具应用到运行和供应链，可持续降低费用，不断创造新的竞争优势和策略，获得更大的效益。

麦肯锡报告指出了零售商使用的 16 种大数据技术，按照价值链走向分为五个主要领域：营销、销售、运营、供应链和新商业模式（James et al., 2011）。

1. 营销

1）交叉销售

交叉销售的最新发展是使用消费者可知的所有数据，包括人口学信息、购买历史、偏好、实时位置以及增加平均销售规模的其他因素。比如，亚马逊网站使用协作过滤方法，在用户访问或购买产品时出现“你或许还想要”的提示。亚马逊曾经一度报告说 30% 的销售额都源自它的推荐工具（Hardy，2011）。这项技术的另一个实例是使用大数据分析对店内优惠促销进行优化，进而推送补充产品或捆绑产品。

2）定位营销

基于位置的营销依赖越来越多人采用智能手机和其他带有个人位置信息的设备。它以接近商店或已在店内的消费者为目标。比如，当一个消费者接近一家服装店，这家店会发送一则特价外套信息到他 / 她的智能手机上。

3）店内行为分析

分析消费者在店内的行为能够帮助提升商店的布局设计、产品组合以及货架摆放。最近的创新让店家可以从智能手机应用程序、购物车应答机或是检测

店内手机的方位来获取实时位置数据，追踪消费者的购物模式（比如在店内不同地段花费的时间和步径）。一些零售商使用和监控摄像头相连的图像分析软件来记录店内行走轨迹和消费者行为。

4）消费者微细分

另一项大数据技术是消费者的微细分。虽然这个概念已经存在，但是大数据带来了巨大的创新。细分所使用的数据量急剧增长，分析工具日益成熟，使得部门还可以进一步细分——直到零售商可以开展个性化服务，而不是简单的市场细分。除了使用传统的市场调查数据和历史购买数据，商家现在可以追踪和使用个体消费者的行为，包括网站的点击流。商家可以将日益精细的数据继续升级到实时数据，以根据消费者变化做出调整。尼曼·马库斯——一家高端商场——采用了行为细分和多层式会员奖励计划，这个组合大大提高了高利润商品在高端客户中的销售额。

5）情感分析

情感分析使用大量来自各类社交媒体消费者的数据流，为多种商业决策提供参考。比如，商家可以使用情感分析测量消费者对营销活动的实时反应从而做出相应的调整。不断演变的社交媒体数据分析在其中发挥重要作用，因为消费者依赖同伴的喜好和评价做出购买决定。现已出现了各种工具可以实现实时监测和反馈网络消费者的行为与选择。

6）提高消费者多渠道的体验

提高消费者的多渠道体验能够有力提升销售额、消费者满意度和忠诚度。商家使用大数据将促销和定价无缝接合，无论消费者是在虚拟商店、实体商店购物还是仅仅在阅读产品目录。比如威廉斯·索纳玛公司将消费者数据与6000 万家庭信息相整合，追踪他们的家庭收入、房屋价格和子女数目。有针对性的电子邮件广告收到的反馈是普通邮件的 10 到 18 倍，公司因此能够制作不同版本的目录，符合不同消费者群体的偏好和行为特征。

2. 销售

1）分类优化

分类优化指根据人口学特征、购买者认知和其他大数据信息来判断哪些产品适合在哪家商店出售，可以极大地促进销售额。比如，一家零售药店使用消费者调查、市场和竞争性分析及详细的经济模型，目的是识别其在产品层面增长乏力的原因。它将总体库存单位减少了 17%，将贴有零售商标签的产品比率从 10% 提高到 14%，实现了 3% 的收入增长以及 2% 的销售额增长。

2）价格优化

当下，零售商家能够使用更高层级的分析工具将定价优化提至新的高度，

可以使用多种来源的数据，近乎实时地对定价决策做出评估，提供参考信息；通过复杂的弹性需求模型观察历史销售数据，了解在库存单位层面的定价，包括减价和调度。商家可以使用数据分析促销活动，评估销售额增加的原因及其成本。一家食品零售店观察到了不同产品目录在不同消费者中的价格弹性。比如，农村食品消费者认为油和米是更高优先级的购物产品，因此这些商品的价格弹性或许低于城市消费者，而后者倾向于将谷类食品和糖果列为优先采购物品。

3）放置和设计优化

实体商店能够通过挖掘库存单位的销售数据、优化物品放置和视觉设计获得极大的增收，本质上，根据足迹信息使用本地化的方式优化设计。在线商店可以通过观察页面互动的数据（比如滚动、点击和悬浮）来调整网页设计。比如，易趣对网页的不同部分进行了几千项试验，以决定最优的布局和其他页面特性，从页面导航到图片规格的大小。

3. 运营

1）绩效透明

零售商家可以每日分析绩效、店铺销售、库存单元销售及每位员工的销售额。如今，这些分析系统更加趋于实时。商家能够通过收款台观察准确度、每小时交易，以及客服质量（顾客投诉以及调查满意度）。虽然这个行业已经广泛采用基础绩效汇报，仍有更加频繁、迅速和颗粒化的趋势，允许管理者对运行情况做出更加及时和具体的调整。

2）劳动投入优化

改善运营的另一个技术是劳动投入优化、自动的时间考勤记录和更好的劳动调度。这项技术可以更准确地预测对员工的需求，特别在高峰期，从而避免生产能力问题。因为店铺劳动力大约代表了商家平均固定成本的30%，使用这项技术将会十分有意义。

4. 供应链

1）存货管理

对多个数据库进行数据挖掘高级分析将会提供更多的细节信息，大数据能够进一步改善商家存货管理。最佳的存货管理在库存层面提供充分的透明度，和自动补货过程相连的条形码系统则减少了销售一空的失误的可能性。主要的零售商通过合并多个数据库（比如销售历史、天气预报和季节性销售周期）来提高存货预测能力。更好的存货管理可以降低商家的囤货标准，因为供应商和需求信号联系更加密切，还会减少因为库存中断造成的销售损失。

2）配送和物流优化

主要零售商也在使用有 GPS 功能的大数据远程信息处理系统（比如远程方位报告）对运输进行优化处理，并使用路径优化功能提高车队和配送管理水平。交通分析法可优化燃料效率，实施预防性维护，督促司机规范行为和优化行车路径。

3）信息供应商协商

在大数据的世界中，主要零售商能够分析消费者偏好和购买行为，从而帮助与供应商进行谈判。他们可以用价格数据和交易数据，将协商的特许权集中在关键产品上。考虑到销售商品的费用占据了最大份额的零售商店费用，此类的大数据应用将带来重大益处。

2.4.2　制造业

制造行业是大数据早期和重度使用者，在电脑诞生之日就开始使用信息技术和自动化技术来设计、制造和配送产品，目的是提高产品质量和性能。在 20 世纪 90 年代，制造业公司获得了惊人的年度生产能力增长，因为运行的改进提升了制造过程的效率，也提高了制造产品的质量。制造商还优化了全球运行和管理，将产品外包给成本更加低廉的地区。相对于绝大多数行业，制造业相对已是非常高效，但是大数据仍然能够提供另一波重大的制造业升级（James et al., 2011）。

1. 产品设计的研究和开发

大数据的使用将会加速产品的开发，帮助设计人员回到最重要和最有价值的产品特性——其基础是具体的消费者投入和减少生产费用的设计，利用消费者的远见，通过公开创新的方式减少研发成本。

1）产品生命周期管理

在过去十年中，制造企业为了管理产品生命周期而采用了 IT 系统，包括电脑辅助的设计、工程、制造、产品开发管理工具和数字制造。然而这些系统生成的大数据集总是受限于它们各自的系统之内。制造商如果建立起产品生命周期管理平台 PLM（Product Lifecycle Management），将多种系统的数据集整合在一起，让有效和一致的合作成为可能，将会抓住非常重要的大数据技术创造更多的价值。比如，PLM 可以为“共同创造”提供平台，将外部和内部的投入综合起来创造新产品。这在航空航天行业将会格外有用，因为该领域的产品往往由全世界数百家供应商提供的成千上万个零件组装起来。在此情况中，原始设备制造商和供应商一同进行设计将会有巨大的价值。PLM 平台还能大大帮助设计阶段的试验。设计师和制造工程师能够以快速低廉的方式分享数据，

建立模拟条件测试不同的设计方案、部件和供应商的选择，以及相关的制造成本。因为设计阶段做出的决定往往占据制造费用的 80%，这样的做法非常有意义。

高级制造行业的主要公司已经开始运用数据和控制实验的协作使用。丰田、菲亚特和尼桑都将新模型开发时间削减了 30%~50%；丰田声称在建立起第一个实体模型之前已经减少了 80% 的缺陷。

2）评估设计

通过市场调研获得消费者的投入和贡献是产品设计过程的常见组成部分，但是很多制造商还没有实现从越来越多的消费者数据中系统地提取出关键的建议，改善已有设计，形成新模型和产品变种的技术标准。最优秀的制造商进行联合分析，试图发现消费者在多大程度上愿意为某些产品特性付费，理解哪些特性对于市场成功至关重要。此外，这些公司还从销售终端的数据和客户反馈挖掘额外的量化的消费者意见。制造商开始挖掘的新数据来源包括社交媒体上的消费者点评，还有描述实际产品使用的传感器数据。

3）开放创新

产品研发和产品创新以回应新的客户需求，制造商们越来越依赖通过新渠道获得外部投入。随着 Web 2.0 的到来，一些制造商开始邀请外部利益攸关方提出创新的想法甚至通过网络平台共同研发产品。生活消费品制造公司（比如卡夫和宝洁）经常征求消费者的意见，并与外部专家合作，包括学界和业界的研究者。进入 21 世纪，宝洁面临研发成本上涨和回报降低的难题。作为回应，宝洁公司设立了开放创新项目，使用 InnoCentive（创新中心）——一个基于网络的平台，公开征集专家对公司面临的技术困难提出解决方案。如今，超过半数的新产品都有来自公司外部的设计因素。宝洁的研发生产率高达 60%，而研发的收益份额从 4.8% 降至 3.4%。这些开放创新项目的确非常成功，但是一个关键问题在于如何将真正有效的设想从大量建议中有效地提取出来。这项使命可以由大数据技术帮助解决，比如自动算法。

通过大数据进行的开放创新还可以延伸到更加高端的产业。比如宝马公司创建了“创意管理系统”以帮助评估那些来自“虚拟创新机构”的创意。这将识别高潜力创意的时间缩短了一半，也减少了评估创意可行性的决策时间。结果是公司每年从开放创意平台甄选出二至三个设计进入其新品模型中。这种创新方法额外的收益是在这些创意活动的参与者中形成更高的品牌效应，以及让这些创新更加广为人知的光环效应。需求的易变性是制造商需要解决的关键问题。零售商客户努力迫使供应商增加弹性和反应性，其原因在于消费者分散的和不断变化的偏好。其他的趋势，诸如采用促销和定价策略，只会让供应商面

对更严重的易变性。

制造商可以充分利用自己的数据提高对于需求的预测以及供给的规划。但是，如同在其他领域显示的，当公司能够将其他来源的数据整合在一起——包括零售商数据，比如促销数据、产品投放数据和库存数据，将会释放出更大的价值。通过应用整条价值链上的数据，制造商可以让大起大落的订单模式变得平缓。这样做的益处可以蔓延到价值链的上下端，实现更有效的现金使用，提供更高水平的服务。最优秀的商家还能加快规划周期的频率，使它们和生产周期保持同步。事实上，一些商家已经开始采用实时数据调整生产量。其他则与零售商协作，使用限时折扣，在店铺层面调整需求量。

2. 生产

可以提高生产效率的大数据工具将虚拟技术应用到在生产过程中生成的海量数据。物联网的普及也帮助制造商使用实时的传感器数据来追踪部件，检测机械装置，指导实际操作。

1）虚拟数字工厂

制造商可以从产品研发和历史上的生产数据（比如订单数据、机器性能数据）获取有用的信息，使用更为先进的计算机方法为整个制造过程建立数字模型。这样的虚拟“数字工厂”包括了所有机器、人工、固定装置，能够用来设计和模拟效率最高的生产系统，从工厂布局到特定产品的生产步骤排序。主要的汽车生产商已经开始使用这项技术优化他们新厂房的生产配置布局，特别是当比如空间和配机设备存在许多限制的时候。炼钢厂可以使用模拟程序为整个资产组合建模，迅速检测出改进方法，这可以将交付可靠性提升 20%~30%。汽车制造、航空航天和国防制造业的案例研究显示，这些先进的模拟程序能够将生产图的变动以及工具设计和建设费用降到最低。

2）传感器驱动的运营

物联网应用的大量增加让制造商可以嵌入来自供给链和生产过程中联网传感器的高度颗粒化的实时数据，由此优化公司的运行。这些数据让全面的过程检测和优化成为可能，减少浪费，将产量或吞吐量最大化。它们甚至可以实现一些迄今尚不可能的制造业创新，包括纳米制造技术。

使用来自传感器网络的大数据的最佳范例来自流程制造业，比如石油冶炼。数十年来，石油产业一直使用大量实时数据以追求始终难以实现的沉淀物。现在该行业将大数据应用到生产方面，对油田进行自动、远程监控。这个方法的好处是削减运行和维护成本（可占浪费开支的 60%）。在数字油田中，单一系统便可以从吸油井管流检测器、地震传感器、卫星遥测系统获得数据。这些数据会被传输到非常大的数据库，转而到检测和调整参量的实时运行中心，优

化产量，缩短故障时间。经验显示，数字油田能够减少 10%~25% 的运行成本，还能提高 5% 或更多的生产能力。

3. 营销和销售 / 售后服务

制造业公司使用来自客户反馈的数据，不仅为了提高营销和销售，也为了做出更加明智的产品研发决策。将传感器植入产品的技术在经济上越来越可行，将会产生大量关于产品实际应用和效能的数据。由此，制造商可以获得关于产品缺陷的实时数据，迅速对生产过程做出调整。进行产品研发时可以应用这些数据进行重新设计和新产品开发。许多建设设备制造商已经将传感器嵌入他们的产品，提供实时数据了解实际使用和使用模式，让制造商能够改善需求预测以及未来的产品开发。

还有使用大数据提高营销、销售和售后服务的机遇。如同许多部门已经实现的，这些机遇从消费者细分到使用分析工具提高销售人员的效益。一种重要性日益凸显的应用是，用来自实际使用的传感器数据提高服务质量。比如，分析来自复杂产品内置传感器的数据，可以让飞机、电梯、数据中心处理器的制造商开发出智能预防性维护服务套餐。这样维修技工甚至可以在用户发现一个部件失灵之前就被派遣去处理问题。

4. 管理

海量数据扩大了算法和以机器为媒介分析的运筹领域。例如，在部分制造企业，算法对生产线的传感器信息进行分析，形成了自我调节的流程，从而减少了浪费，避免了代价高昂（有时十分危险）的人为干预，最终提升产量。在先进的“数码化”油田，仪表不时读取有关井口状况、管道和机械系统的各类数据。这些信息由一组计算机进行分析，并将结果输入实时运营中心。后者则调整油量以优化生产和最大限度地缩短停机时间。一家大型石油公司因此减少了 10%~25% 的运营成本和员工成本，产量提高了 5%。

大数据时代还可以形成新的管理原则。在专业化管理的早期，企业领导人发现最小有效规模是成功的关键决定因素。同样，对于能够捕捉更多更好的数据，而且还能够高效化、规模化利用它们的企业来说，竞争优势将不期而至。

2.5　物联网

物联网（Internet of Things, IOT）的概念作为下一次信息技术升级的关键词而广为人知，它是以信息感知为特征的物联网，被称为世界信息产业的第三次革命，目标是将现有的、虚拟的互联网拓展到现实世界，使得任何真实世界的对象都可以自动加入网络，从而在全球范围内实现追踪和查询。

物联网涵盖了多种技术和研究领域——诸如自动识别、无线传输、综合传感、分布式数据处理等。温家宝总理在十一届全国人大三次会议上作的《政府工作报告》中对物联网的定义是:物联网是通过信息传感设备,按照约定的协议,把任何物品与互联网连接起来,进行信息交换和通讯,以实现智能识别、定位、跟踪、监控和管理的一种网络。它是在互联网基础上延伸和扩展的网络。

物联网所需的体系结构分别为:射频识别、传感器技术、嵌入式逻辑对象、对象的特设网络、基于互联网的信息基础设施。

实现物联网的潜在好处是多方面的,既为个人也为企业。一些最有希望的应用包括:改善全球供应链物流的管理、假冒产品检测、生产制造自动化、智能家居和家电、电子政务(电子公文和货币),以及电子医疗(病人监测和病人记录)(Lopez et al., 2012)。

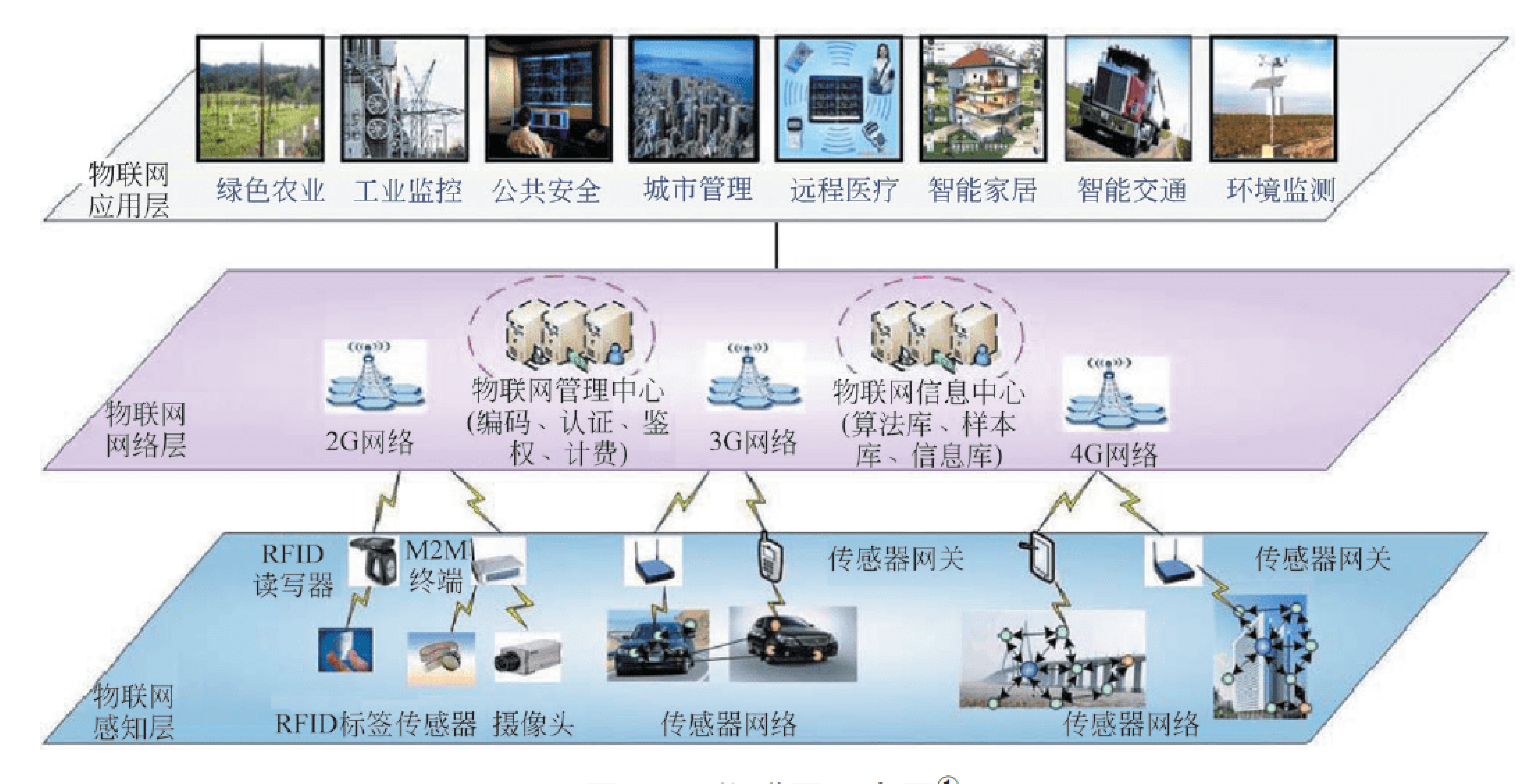


图2-9 物联网示意图[①]

物联网已经成为我国的战略性新兴产业。通过物联网可在传统工业、生产安全、工程控制、交通管理、城市管理、农牧林业生产、商业流通等多领域建立随时能在物体和物体之间沟通的智能系统,有利于推进信息化的进程,并对我国的各种产业产生重要的影响。

2.5.1 物联网的基础

1. 成熟的传感技术

随着微电子技术的发展,涉及人类生活、生产、管理等方方面面的各种

① 资料来源:中国移动物联网实验室。网址:http://www.iot.10086.cn/。

传感器已经比较成熟，例如常见的无线传感器（WSN）、无线射频识别（Radio frequency identification devices）、电子标签等。其中，无线传感器可以提供关于多种环境特点的持续数据流，注入物联网。其他更加高级的设备包括通过传感设备识别物体，比如通过电子图像处理技巧。个人计算机和智能材料的使用大大提高了人类和物理环境之间的交互性。生物识别技术能够用来实现物联网的安全性和个人化。

2. 宽带互联网络

网络发展到今天，已经真正进入到"信息高速公路"时代，使得我们可以以各种方式接入到网络，如光纤、宽带、WIFI、无线个域网（Zigbee）等。

3. 高速信息处理能力

计算机软件技术迅猛发展，计算机的存储能力、计算能力还在进一步增强，基于海量信息收集和分类处理的能力大大提高。

2.5.2　物联网的数据类型

物联网应当实现对信息的获取和精确控制，而绝非信息的大量耗散结构。通常来看，物联网所需的数据分为以下五类：识别数据、位置数据、环境数据、历史数据和描述性数据（Cooper et al., 2009）。

1. 射频识别

射频识别指的是使用无线电波进行物品的识别和追踪，这项技术正在变得寻常可见。RFID（无线射频识别）的标签可以置入物品中，用来传递和接受信息。RFID 源自二战时期的技术革新。最早的商业应用始自 20 世纪 80 年代，标准在 20 世纪 90 年代出现，更广泛的应用一直延至今日，成为日常生活的一部分。

2. 地址 / 唯一识别符

物联网的物体需要用唯一的 IP 地址识别。随着物联网的范围扩大，所需的识别符将会增长。正是由于物联网的应用，IPv6 才被选为 IPv4 的替代物——后者的 IP 地址正在耗尽，而前者使用 126 比特的地址，容量远远大于只使用 32 比特的 IPv4。

通过多级层次命名的方法可以提高识别能力。全球各地的局域识别符都受到域名的限制，就像目前互联网中域名的使用方式那样。互联网已经开发出完善的命名方法，由互联网地址编码分配机构（Internet Assigned Numbers Authority）审查全球的 IP 地址分配、跟区管理、域名以及其他网络协议相关的任务。

3. 关于物品、过程和系统的描述性数据

物联网的价值主要来自网络中的物体、过程和系统所记录的数据或是元数据。元数据是关于数据的数据，对于使用者寻找和获得合适的数据至关重要。我们能够这样区分二者：举例来说，物品“24.672.673.982”数据是猫、黑色、有绒毛的，相对应的元数据则是类型、颜色和姓名。数据的存储、表现、验证，如何保证数据检索和更新的最大效率和不可否认性，都是研究者关心的话题。物品需要能够自描述，并且能够报告动态特征，以便于数据最大化的分享。

需要存储的不只是基本物品数据，还有过程数据和系统数据。系统和过程可以被视为特殊种类的物品，属性更加复杂。存储物品、过程和系统的数据十分重要，这样用户才能掌握如何利用物联网所提供的服务和设备。比如，在一户住宅中，网络可以收集一段时间内用电量的环境数据，这是关于物品的数据。过程数据则是计算该段时间内用电量的峰值和低位。然而，这样的一个过程或服务可能只是和物联网所提供的成千上万个其他过程一样，难以定位。这种情况下，关于过程和自描述数据的元数据以及标引系统就很有用。

4. 定位数据和无处不在的环境数据

定位数据能为一个特定标签物品提供位置信息，无论在全球定位系统中（GPS）或是地方定位系统中。GPS 依靠多个卫星将信号传递给调节单元，物品可以通过三角测量确定自己的位置。地方定位系统有类似的运行机制，只是覆盖面更小。地方技术的例子是蜂窝式基站、无线网访问点和电视信号塔。地方定位系统可以与 GPS 合作，有时还可以替代后者。它们可以用于建筑或楼宇密集区。定位数据组成可静可动，它将会在物联网中发挥重要作用。互联网中一种新型信息是普遍的定位信息。它是关于环境的全部信息，并不显眼，但是能够改善和帮助我们与周围环境的相互作用。这个信息也依赖于定位。现在已有的不仅是物联网的概念，还有 internet of place——所有关于某个地点的特定信息都能被该地区的设备和用户迅速获得。最终所有的地方都被纳入这个网络。与此相关的技术包括移动计算、地理信息系统以及环境技术。

5. 传感器数据

数据进入物联网的路径之一是通过无线传感网络（WSN）。电子技术的进步使得建立一个 WSN 变得相对容易，以此可用来检测各种各样的环境现象，比如天气、温度和噪音。个域网联盟（Zigbee Alliance）已经生产出标准支持 WSN 的装配。这项技术中涌现了不同的、有趣的研究方面。比如，数据收集是持续性的、间隔性的，还是仅在查询时才进行？其他的问题还有，我们如何保证有效率地获得具有代表性的样本，我们应该将多少数据存档？传感器和网

络技术让迅速抓取海量数据成为可能，但是查询并且挖掘数据可能出现问题，特别是当需要进行实时的分析。目前已经提出一些技术来解决以上的问题。

6. 历史数据

物联网的传感器可以收集拍字节甚至更大量的数据。这些数据可能需要进行存储。随着时间流逝，这些数据变成了历史数据。数据的体量成为了难题。需要做出应用为导向的设计：尔后如何保有数据？哪些数据应该保留？一些将会保留在活跃的数据货栈中，便于频繁的查询，另一些的需求量较小，可以存储在不太容易获得的地方。对数据存储的问题已经引起注意，比如数据丢失，不准确的记录，缺失信息以及对废弃技术的依赖。数据库社区提供了一些解决方案以实现更加有效的数据存储。这些方案也适用于物联网。

7. 物理模型

物联网的应用需要物理模型，以便能够在运算法则中使用。物理模型是现实世界的模板，比如重力、力量、声音和磁性。呈现这些模式可以提供建模和物理场景的模拟物。物理模型目前在电脑游戏和电脑辅助的工程领域内十分普及，将它们合并进物联网将会提高它的功能。

8. 监控所用的执行和命令数据

物联网将会被用来远程控制设备，这需要设备制动器状态的反馈。特别是因为相关应用的实时性特质，如何表现制动器状态成为了一个挑战。为了支持互联网制动器越来越多的使用，需要的技术革新是微型化和节能电子产品，后者包括低耗微型计算机和传播方法、能量采集转换器、改进的微型电池。

一些进入物联网的数据是控制设备的指令数据。比如，一个人可能突然在半个小时之内回到家中，希望打开暖气做好准备。用户需要在物联网中控制设备，这则需要一种特殊的语言。物联网中的不同系统可能开发出不同产品，并具备不同的源头，因此并不存在同样的指令界面。需要努力将指令 / 控制数据和界面进行标准化操作。

2.5.3　物联网的最新发展和应用

在物联网的实际应用层面，个人定位数据是至关重要的概念，是信息数据系统的基础。它指的是一个人或一部设备的实时位置，通常表现为一个数字编码，可以在覆盖全球的坐标网（Grid）上标记出个体的所在位置。最早的个人定位数据来自信用卡和借记卡的支付信息，它与刷卡所用的 POS 终端（通常是在固定的地点）显示的个人身份信息相连。相似的数据来源是发生交易的 ATM。2008 年，全球范围内和 POS 设备相连的离线交易为 9000 万次到 1 亿

次（James et al., 2011）。

随着手机使用者数量的增加，使用基站信号对这些设备的位置进行三角测量变得愈发普遍。这项技术能够识别将近 50 亿使用者的方位。智能手机的使用也在增加。在 2010 年，同时在线使用手机的人数达 6 亿，这个数字预计将会有 20% 的年增长率。智能手机带有 GPS 和无线上网的功能，这两项技术都可以确定手机的所在位置，使得个人定位数据更加准确、更容易获得，特别对于手机实用程序的开发者而言。

根据麦肯锡全球研究所的报告，2009 年全球范围内的个人定位数据总量达到 1 拍字节（相当于 2 的 50 次方），并以 20% 的速度增长 。装有 GPS 功能的智能手机的爆炸性增长是这个增长的首要推动力。值得注意的是，相对于医疗卫生领域所用的图像或是视频，确定方位的数据量不过几个字节，这意味着每个字节产生的价值远远高于前者。

1. 个人层面的应用

1）智能路线选择

基于实时交通信息的智能路径选择是个人定位数据使用率最高的实际应用。更加先进的导航系统能够实时获得交通信息，包括事故、道路施工和拥堵地段。这样的导航系统还能为用户提供最新的个人兴趣信息和天气状况。职能路线选择设备不仅为驾驶人提出避开拥堵地点的建议，还能将位置和行驶信息传回一个中心服务器，更加准确地计算拥堵的程度。

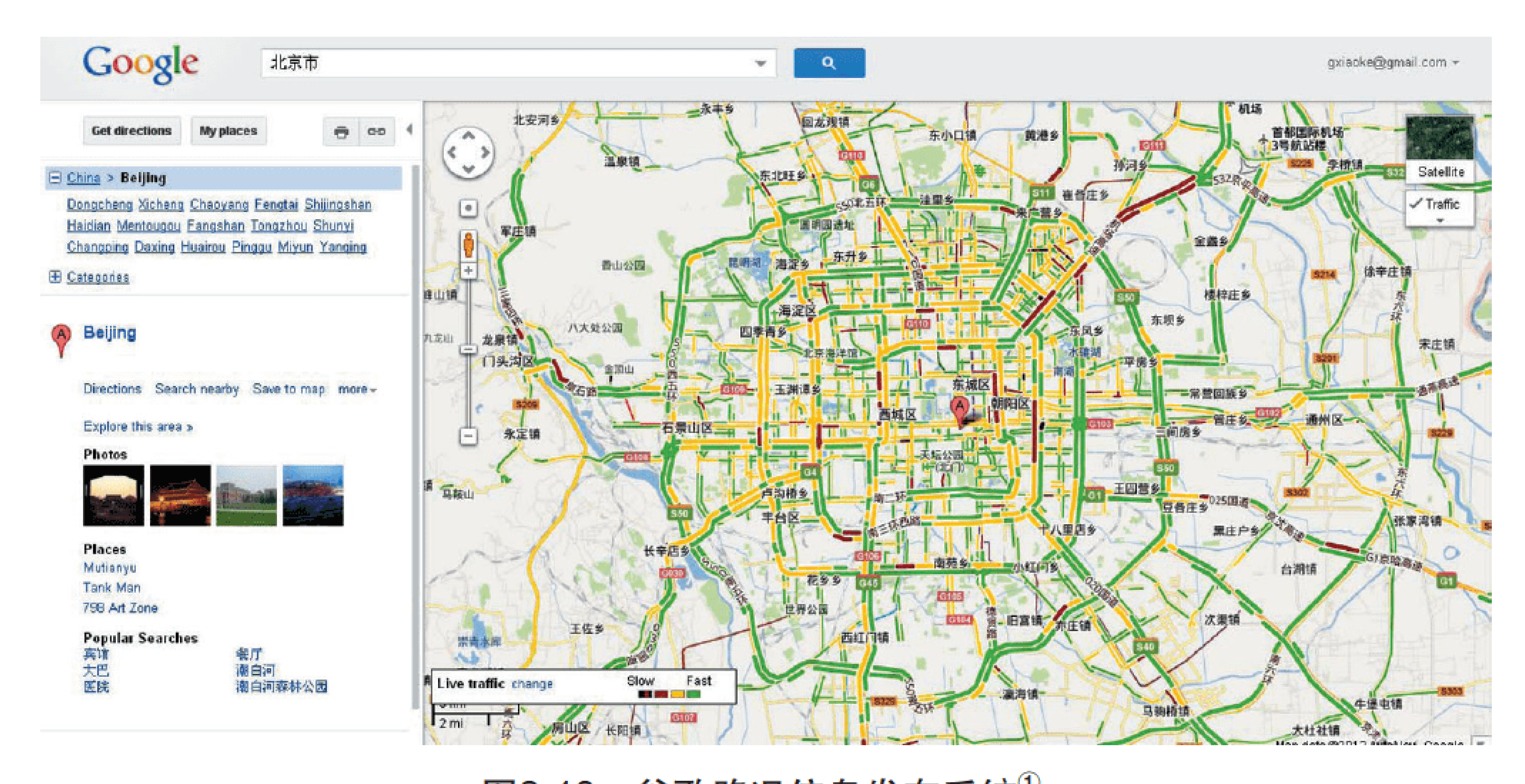


图2-10 谷歌路况信息发布系统[①]

① 资料来源：http://www.ditu.Google.com，谷歌地图能实时反映路面交通拥堵状况。

麦肯锡预测，到 2020 年，智能路径选择的全球价值（以时间和燃料节省为形式）将达 5000 亿美元，这相当于为司机们节省 200 亿小时，或是每个旅行者每年节省 10~15 小时以及 1500 亿美元的油耗。这些结余可以转化为减少 3.9 亿吨碳排放。这些预测数字意味着数字地图和实时交通信息这两项技术的充分开发。目前，大多数发达国家已拥有数字地图数据，正在向发展中国家普及。

2）汽车通讯系统

物联网在汽车行业中的应用包括了使用“智能物体”进行全面检测和报告——从胎压到和其他车辆之间的距离。RFID 技术可以简化汽车的生产，改善物流，增强质量管理，提高售后服务。汽车的每个部件都含有制造商的信息、生产时间、地点、序列号、型号、产品编码等信息，甚至可以显示它们在车辆中的具体位置。专用短程通信（Dedicated Short Range Communication）① 还可以实现更高的比特率，降低和其他仪器相互干扰的可能性。汽车对汽车（V2V）以及汽车对基础设施（V2I）通讯将会极大提升智能交通系统（ITS）的应用性能，比如车辆安全和交通治理（Vermesan et al., 2009）。

在未来几年内，越来越多的车辆将会配置 GPS 和远程信息处理系统，提供一系列个人安全和监控服务。一项已有的例子是，通用电气的 OnStar 服务能传送实时车辆位置和诊断信息给中心监控点。这类似医疗系统中的远程健康监测，能够在设备需要修理或软件升级的时候提醒司机，或是在紧急情况下为车辆定位。

3）手机定位服务

手机定位服务是一项快速发展的技术，拓展了其他由手机提供的定位服务（LBS）的范围，比如追踪儿童和家属的安全应用程序。已有的例子包括 Foursquare（2011 年 4 月用户已达 800 万人）和 Loopt（2011 年 4 月用户超过 500 万人）。Loopt 于 2006 年成立于加州山景城，是一个智能手机下载使用的应用程序，允许使用者在好友之间分享实时位置信息、状态消息和带有地理标签的照片。现在美国所有主要社交网站都提供详细地图信息，可以显示好友的位置、正在做什么以及如何找到他们。Loopt 的主要盈利来自地理标识的广告和促销信息。

类似定位服务的收益模式都将是免费服务和应用（广告商付费）及其他收入的组合，包括赞助商的链接（餐馆、酒吧和其他兴趣点）。

① 专用短程通信是在车辆与路边设备之间进行的无线通信，也就是运输子系统中的车辆子系统与道路子系统间的通信。



图2-11 手机定位服务[②]

2. 组织层面的应用

1）地理定位广告

地理定位的手机广告是个人位置数据获得价值的最常见方式之一。例如，消费者可以选择接受某个地理定位广告，一旦接近某个喜爱的商店，个性化广告就会出现在智能手机上。在餐厅里就餐的广告用户可能会接收到该餐厅发来的优惠券。这项技术可以基于智能手机用户的方位或目的地，提供关于最近的ATM、餐厅点评，或是商店促销的信息。

和电视、平面广告这些更传统的广告方式相比，地理定位广告对消费者的购买决定有更大的影响，可以提高销售量。例如，ShopAlerts 由位于旧金山和纽约的 Placecast 公司研发，是一款定位“推送短信”产品，目前已有包括星巴克、REI、AE 服装、Northface 等公司使用该产品为自己揽客，在全球范围内用户达到 100 万人。这款产品在美国可以为超过 90% 的手机定位，该公司报告称 49% 的访问量是在顾客收到 Shopalerts 的短信后发生的，而另外 19% 的短信起到了提示作用（Sims，2012）。

2）电子收费

目前的电子收费产品所需要的专门技术成本十分昂贵，但是配备 GPS 功能的手机的普及将会刺激对收费设备的应用研发，降低该系统的成本。举例来说，一部手机可以为车辆和收费站定位，使用用户的手机账户支付费用，不再需要单独的应答机和付款账户。

3）保险定价

个人定位数据和汽车信息处理器有可能为保险公司提供更加准确和详细

② 资料来源：http://labs.chinamobile.com/news/53955_p.4。

的个人行为信息——比如投保人的驾驶习惯。这些信息让保险公司根据个人的实际行为定价，而不是参考笼统的人口统计学。一些人认为基于行为的保险可以减少索赔支出，因为当人们得知自己的行为受到监控，其行为会更加谨慎。这个结论有待于验证，但有一点可以肯定：根据个人定位数据而研发的技术能够帮助保险公司设计出鼓励安全驾驶的服务。比如，保险公司可以提供关于天气、交通情况、高危停车场以及道路限速的实时警告。

4）紧急响应

个人定位数据、实时交通信息以及 GPS 数据传输系统可以让执法部门、火警和急救车更快、更有效地执行任务。这些技术让紧急行动的调度员能够迅速识别突发事件报告人的所在地，保证行动组可以尽快响应（通过智能路径选择），并保证他们在危险环境中的自身安全。

5）城市规划

个人定位数据的集合分析能够在宏观层面帮助决策，这主要体现在两个迥异的领域：城市规划、智能城市领域和经济、商业领域。

对个人定位数据的分析能够极大地帮助城市规划师的工作。通过分析道路和大容量公交运输建设、交通拥堵的缓解方案、高密度开发区的设计，规划师的决策判断能力将会得到提升。城市交通和发展规划师将会越来越多地使用海量信息——比如高峰和非高峰期交通拥挤地段的流量与模式、购物趋势——来减少拥堵和污染物的排放（Ibarguen et al., 2009）。城市规划师将会深度挖掘和观察大量数据，为决策提供充分翔实的背景知识，从而做出更加明智的选择——从交通信号灯的摆放位置到停车空间的需求。新加坡的公共交通部已经开始使用十年期需求预测（部分基于个人定位数据）来规划交通需求。荷兰的交通局则使用来自手机的个人定位数据预测交通和行人的拥堵情况。

2.5.4 发展前景展望

15 年前，美国最庞大的数据仓库的规模只有数 TB，且只有像沃尔玛、万事达这些极少数的公司才拥有如此庞大的数据。而如今，从生活中的购物交易到工业上的生产制造，从社交网络媒体信息到在线视频图像资料，从企业的信息管理系统到政府部门的电子政务，都有着大量的数据产生。当 TB 骤然剧增到 PB（1PB=1024TB），常规技术显然难以应对需求。

2012 年 4 月 11 日，中国英特尔物联技术研究院正式成立。计划在未来 5 年里共同投资 2 亿元人民币，着力攻克智能感知、传输技术、大数据处理技术和共性技术基础研究等核心物联网技术。4 月 12 日，英特尔科学技术中心首席研究员 Mei Chen 和英特尔中国研究院首席工程师吴甘沙发表了主题为“在

物联网时代驯服大数据”的演讲[①]。

中国工程院院士邬贺铨断言，物联网所带来的产业价值要比互联网大 30 倍，物联网将形成下一个上万亿元规模的高科技市场。这个数据来自于美国权威咨询机构弗雷斯特（Forrester），后者预测说，到 2020 年世界上物物互联的业务跟人与人通信的业务相比，将达到 30：1，因此，“物联网”被称为是下一个万亿级的产业。

赛迪顾问的研究也显示，2010 年，我国物联网产业市场规模达到 2000 亿元，2015 年我国物联网产业整体市场规模将达到 7500 亿元，年复合增长率超过 30%[②]。

（本章编译者：张小娅，清华大学国际传播研究中心助理研究员，博士生）

---

① 详见网易科技报道，http://tech.163.com/12/0411/14/7UQNAJJ600094MLL.html。

② 中国计算机行业网，http://www.ciw.com.cn/newsdemo/ciwnews/201003/20100308110630.shtml。

第3章 大数据挖掘

3.1 引言

早在五千多年前，两河流域的苏美尔人就用泥板和树枝记录纳税信息，中国的甲骨文也向现代人透露来自远古的故事。“记录”这一人类独有的行为的重大意义在于，人类创造的历史和文明可以被继承和延续下来，后代的人能够循着前人的足迹继续前进，“记录”是联系历史与现在的纽带。当下，人类活动正在以各种方式被记录着，从出生证明到死亡证明，从逛街购物到生病住院无所不包。记录的载体从泥板到羊皮，从纸张到电脑硬盘，这些记录的工具和形式随着人类文明的发展而发展，反之，这些记录也促进着文明的进步。

21 世纪的人类已经有能力存储大量的记录，据测算，目前全世界的数据量相当于地球上每个人拥有一块 100G 的硬盘。这些被记录和存储的内容是如此的庞大和复杂，如何利用它们正日益成为重要的议题。当下，大数据挖掘的需求应运而生。“数据挖掘”（Data mining）这一词汇本身就暗示着从看似平淡无奇的记录资料中发现、归纳和获取有价值信息的过程。现代人从大数据里挖掘价值的过程与古老的沙里淘金有着惊人的一致性，只不过对象从实物变成了抽象的代码。

数据挖掘①概念的正式产生肇始于 20 世纪 80 年代开始的计算机和信息技术的发展。随着计算机存储容量的增加，许多公司开始储存更多的交易数据。由此产生的记录集合通常被称为数据仓库（Data warehouse），因其太大，无法

① 内容编译自《大英百科全书》。在线“data mining”，*Encyclopedia Britannica. Encyclopedia Britannica Online Academic Edition*. Encyclopedia Britannica Inc., 2012, Web. 18 May. 2012, http://www.britannica.com/EBchecked/topic/1056150/data-mining。

与传统的统计方法进行比较分析。于是科学家们召开了许多计算机科学会议和研讨会，主要议题就是讨论如何利用人工智能领域的最新进展——例如机器学习，神经网络等技术——来进行的数据库知识发现（Knowledge Discovery in Database，简称 KDD，是“数据挖掘”在计算机学科中的首选术语）。而后 1995 年在加拿大蒙特利尔召开了第一届“数据挖掘和知识发现”国际会议，同名刊物也于 1997 年创刊。在此期间，出现了许多早期的数据挖掘公司和产品。

商业出于对利润的敏感，往往能嗅到技术发展带来的赚钱机会，数据挖掘技术与商业情报（Business intelligence）的需求一拍即合。最早成功应用之一便是探测信用卡欺诈。通过搜集用户的刷卡记录，信用卡公司分析了这些记录与持卡人信息特点之间的关系。当时的数据挖掘人员通过分析一段时间内不同类型持卡人的刷卡行为，得到某类持卡人 / 消费者的一个典型的消费模式。于是，当这张信用卡被盗刷，或者是持卡人意图进行信用卡欺诈时，信用卡公司会通过刷卡终端搜集到处于这个模式之外的消费信息，公司可以标记这些持卡人并为后续的调查做准备，甚至拒绝交易。除此之外，信用卡公司掌握的刷卡记录还被用于研究消费者喜欢在什么地点、什么时间，购买什么东西，或者哪些物品会一起被购买，分析人员把重复出现的物品放在一个购物篮子中，得出的结论可以提供给商家，告诉他们：将哪些物品捆绑销售，捆绑打折销量会提高。这些从普通销售记录中“挖掘”出来的有价值的信息被出售给百货公司，以供它们改进营销策略。如此种种的数据利用形式是数据挖掘的一种传统形式，仅仅发掘到了“记录”宝藏的一小部分。

正像由于勘探和开发技术的限制，目前许多海底矿产能力还无法被人类利用一样，目前人类面临的信息的海洋也蕴藏着无与伦比的宝藏而不能为人所用。大数据挖掘技术当前迫切需要发展。大数据时代的数据量远远比“矿藏海洋”这个概念浩瀚得多,而且这些数据还在以惊人的速度和加速度进行增长。另外，数据的类型不仅仅局限于可以抽象的“数字”，还扩展到大量的文本、超链接、音频和视频等传统挖掘方式难以着手分析的信息,如何从中寻找到“金子”是当下和今后信息产业的新发展方向。

本章主要试图向非计算机科学读者介绍大数据时代背景下，大数据挖掘的一些基本知识，包括已有的一些模式和方法以及面临的变化与挑战。本章将介绍目前计算机科学中对于数据挖掘既有的基本路径和思路，继而介绍数据挖掘技术上的一些工具和方法。最后在宏观上讨论数据挖掘对大数据时代意义和挑战。讨论中涉及社交网络的用户行为、网上购物者的消费行为分析以及大众媒体中文本、视频和音频数据挖掘的例子。

3.2　路径和思路

数据（data）的含义可以多种多样。本章所讨论的数据，指的是可以被计算机、互联网服务器及各种终端记录、传输和分析的信息。

当我们的手中掌握了大量数据，我们能拿它们做什么呢？有两种常见的数据挖掘研究路径。第一种，也是比较传统的方式，我们可以称之为“假设检验”的方法。这也是统计学中常用的方法。它指的是在真正利用数据之前，数据挖掘人员脑海之中已经有了一个前因后果的理论假设，他需要利用手中已经有的数据来证明这个假设是否属实。比如说，信用卡公司管理层认为，教育程度越高的持卡人每月平均消费额越大（这就是一个假设）。于是该公司的数据员会从公司的信用卡持卡人数据库中抽出教育程度及其对应的消费记录数据，通过计算机辅助运算看看教育程度和消费额度之间有没有正向相关的关系，当然其中还会涉及很多数学和统计学运算模型，我们这里忽略掉细节。利用数据对已经存在的假设进行检验，是数据挖掘一开始常用的路径。上述简化的例子代表的是从“假设—数据—验证”的过程。

而第二种数据挖掘的路径是数据库知识发现（KDD），也是目前计算机科学中正在积极探索的路径。这种方式中不存在预想好的假设或者论断，而是在掌握大量数据的基础上，通过“观察”数据本身而获得。“观察”似乎是计算机不具备的能力，而海量数据的计算又是人工难以完成的，所以人类需要设计一些具体的程序让计算机学会“观察”，这常常需要借助数据可视化工具或者计算机分析数据中各个因子相互之间的相关程度等方式。这个“数据—结论”的路径是一种更直接的方式。由于技术的限制，在过去前一种数据挖掘路径成为了主导，而近年来呈爆炸式增长的数据迫使人们消除技术壁垒，在第二种路径中寻求突破。从 1997 年开始举办的数据挖掘世界杯（KDD Cup①）就是一个专门针对数据库知识发现的竞赛，它向业界和学界开放，为竞争者提供一个数据库和数据挖掘任务。2012 年竞赛的数据挖掘任务是中国互联网公司腾讯提供的“腾讯微博的社交网络挖掘”和“搜索引擎日志中挖掘用户点击模式”。2012 年 8 月 KDD Cup 的获奖者将在北京同期举办国际数据库知识发现和数据挖掘年会上颁奖。本年的竞赛和年会主题都是围绕大数据的挖掘展开，体现了业界和学界对此的高度关注。

在近年的研究者中，戴维·奥尔森和石勇（2007）介绍了被广泛应用的跨行业数据挖掘标准流程（CRISP-DM），可以让读者了解数据挖掘的一般过程。这个标准流程包括了六个阶段：

---

① 关于比赛具体介绍详见以下网址：http://kdd2012.sigkdd.org/kddcup.shtml。

业务理解：数据挖掘人员确定工作对象、了解现状，制定工作目标和工作计划的过程。

数据理解：一旦对象和工作计划拟订了，就要考虑所需要的数据。这一步骤包括原始数据搜集、数据描述、数据探索和质量核查。这一步骤和第一步常常需要反复进行。

数据准备：就像做菜需要对食材进行筛选、洗净、切成一定形状一样，原始数据中有大量错误、重复的信息，需要删除、整理和转化。数据准备可以视为一次数据探索，为之后的模型建立做准备。

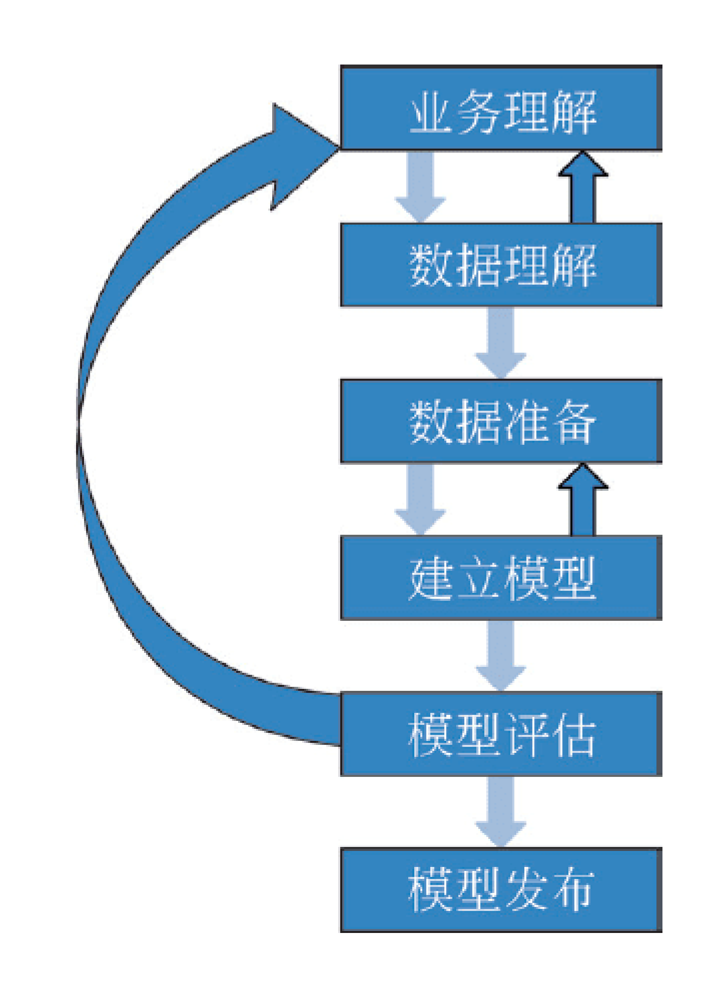


图 3-1 CRISP-DM数据挖掘过程

建立模型：这一阶段需要描绘数据并建立关联，然后用一定的分析方法借助数据挖掘工具进行数据的基础分析。

模型评估：模型结果要对在第一步建立的工作目标进行评估，这将导致频繁地返回到前面的步骤。这是一个缓慢推进的过程，各种可视化分析结果、统计和人工智能工具将向数据挖掘人员展现更深层次地理解数据运行的关系。

模型发布：数据挖掘应用于先前提到的两种路径中，借助 CRISP-DM 前期步骤中发现的知识，可以获得更加健全的模型。这个模型可以用于预测或识别关键特征，需要在实际情况下检测其变化。如果发生重大变化，模型就需要被重新制定。模型发布就让从实验数据库中建立起来的模型在实践中受到检验。

Pyle（1999）在被广泛引用的《数据挖掘中的数据准备》一书中强调了数据挖掘前准备工作的重要性（见表 3-1）。他把数据挖掘的工作总共分为四大部分：探究问题、探究解决方案、特定工具选择、数据挖掘。前三部分工作占用的时间占总时间的 20%，在重要性上却占到关键的 80%。这个划分方法虽与 CRISP-DM 的六步模型不同，但是两者都强调了第一步——思考问题及其相关的方案和选择适合工具的重要性。

表 3-1 数据挖掘不同阶段所占时间和重要性[①]

	占总时间的百分比 /%	合计	对于成功重要性的百分比 /%	合计
1. 探究问题	10	20	15	80
2. 探究解决方案	9		14	
3. 特定工具选择	1		51	

① 资料翻译自：“Figure 1.1 Stages of a Data Exploration Project Showing Importance and Duration of Each Stage.” Data Preparation for Data Mining. San Francisco, CA: Morgan Kaufmann, 1999. Print。

续表

	占总时间的百分比 / %	合计	对于成功重要性的百分比 / %	合计
4. 数据挖掘		80		20
a. 数据准备	60		15	
b. 数据调研	15		3	
c. 数据建模	5		2	

3.3 准备数据

3.3.1 挖掘方法

数据挖掘的方法实际上可以视为在大量前人工作的基础上形成的计算机“思维模式”。奥尔森将数据挖掘的方法分为类别、估计、聚类和概要四个类型。类别和估计都是属于事前预测性质的，而聚类和概要则是事后描述性质的。本节介绍一些最常用的数据挖掘方法，而不能罗列所有的数据挖掘方法。

1. 聚类分析

聚类分析常常是最初的分析工具。它能够在你拿到数据之后对其进行合适的分类。聚类分析是以数据为基础的，它不具有预测性，它的功能是发现数据之间的相似性，并进行分组。在引言中提到的信用卡诈骗检测中，信用卡持有者的信息丰富，既有教育程度又有年收入，既有职业也有性别等。按照怎样的划分对信用卡公司管理和向客户销售是最有效的？聚类分析就是帮助第一步分类的方法。分类的不同往往会导致结论的不一样。一个容易理解的例子就是，目前人类观察到的恒星数量达 10^{12} 个数量级，如何将这些恒星分类？分类的依据可以是与地球的距离、体积、质量、亮度等。在各种各样的分类之中，科学家利用温度和亮度为坐标将这些恒星有效地分类，使得方便天文学家进行研究，得出恒星演化的理论。据此画出的恒星分布图叫赫罗图（见图 3-2），从左上

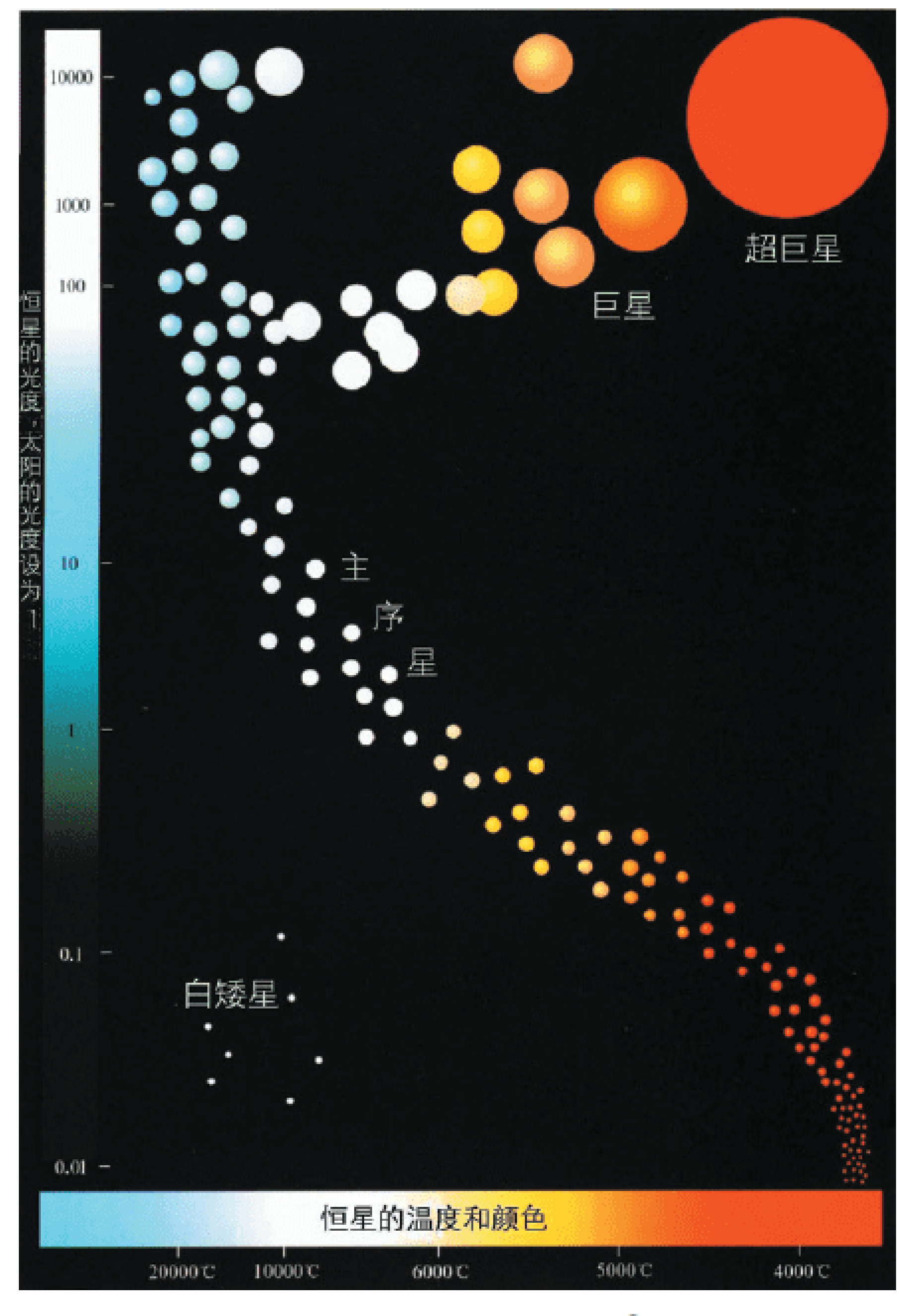


图 3-2　赫罗图[①]

① 图片来自 NASA 官方网站：http://heasarc.gsfc.nasa.gov/docs/RXTE_Live/class.html。

到右下的斜线上分布的恒星是主序星，主序星右上方和左下方分别是红巨星（Red giant）和白矮星（White dwarf）。这两个坐标很好地划分了恒星的类别，成为天文学的一个基础性工作。

2. 回归分析

回归是一个基本的统计学工具。在数据挖掘中它也是一个基础的分析工具，它可以描述一个或几个自变量和一个因变量之间的关系。这种关系可以是线性的也可以是非线性的。通过概率和统计的数学方法，利用手头的自变量和因变量的数据可以找到两者之间对应的数学关系，即得到一个模型，在这个模型中可以利用自变量对因变量数据进行预测。传统的软件都可以进行回归分析，如 SAS、SPSS 或者 Excel。

3. 神经网络

神经网络是受到人类大脑各个神经细胞工作方式的启发，构成的一个网状结构系统。这种网络由一个一个微小的处理器（类似于神经元）和各个处理器之间的弧线（类似于神经线）构成输入层、隐藏层和输出层。神经网络的特点是中间有隐藏层（见图 3-3，从左到右），人们从输入层录入数据，各个微小处理器之间可以模拟类似人类的识别、记忆、思考过程，从而得出结果。神经网络处理数据的优点是有高度平行处理的能力，而且可以有识别、学习能力；此外，出现部分的计算差错或者是数据错误不会影响整体计算过程，就像人类大脑部分受损之后并不影响其整体工作一样。但它也有明显的缺陷，即隐藏层是人们无法解释其运算过程，就像一个黑箱一样。而且多组平行的、部分隐藏的数据通路让人无法判断哪一个通路是最优的。

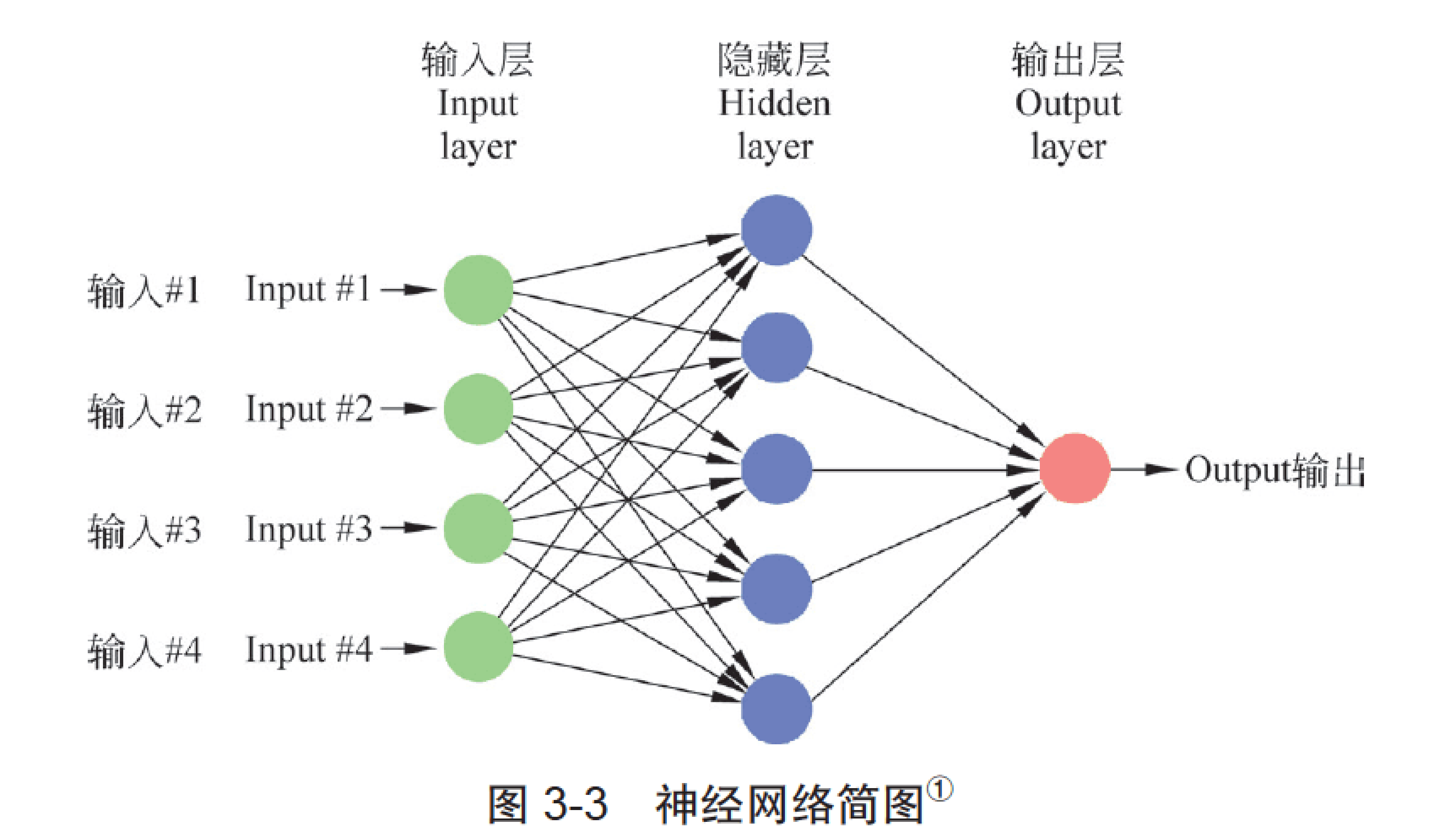


图 3-3 神经网络简图[①]

① 资料翻译自：Fauske, Kjell Magne. *Neural Network. Digital image.* Texample.net. 7 Dec. 2006. Web. 22 May. 2012, http://www.texample.net/tikz/examples/neural-network/。

4. 决策树算法

决策树模型是一个被广泛使用的思考工具，它也是数据挖掘的基本方法。

据韩慧等人（2004）介绍，决策树算法的分类学习过程包括两个阶段：树构造 (Tree Building) 和树剪枝 (Tree Pruning) 。

① 树构造阶段

决策树采用自顶向下的递归方式：从根节点开始在每个节点上按照给定标准选择测试属性，然后按照相应属性的所有可能取值向下建立分枝、划分训练样本，直到一个节点上的所有样本都被划分到同一个类，或者某一节点中的样本数量低于给定值时为止。这一阶段最关键的操作是在树的节点上选择最佳测试属性，该属性可以将训练样本进行最好的划分。选择测试属性的标准有信息增益、信息增益比、基尼指数 (Gini Index) 以及基于距离的划分等。此外，测试属性的取值可以是连续的 (Continuous)，也可以是离散的 (Dis-crete)，而样本的类属性必须是离散的。

② 树剪枝阶段

构造过程得到的并不是最简单、紧凑的决策树，因为许多分数反映的可能是训练数据中的噪声或孤立点。树剪枝过程试图检测和去掉这种分数，以提高对未知数据集进行分类时的准确性。树剪枝主要有先剪枝、后剪枝或两者相结合的方法。树剪枝方法的剪枝标准有最小描述长度原则 (MDL) 和期望错误率最小原则等。前者对决策树进行二进位编码，最佳剪枝树就是编码所需二进位最少的树；后者计算某节点上的子树被剪枝后出现的期望错误率，由此判断是否剪枝。

图 3-4 是一个简单的决策树运算过程。假设在固定存款和购买股票之间选择，假设股票的预期收益有优、中、差三种，可能性分别是 30%、40% 和 30%，与固定存款带来的固定收益比较起来哪个更好？通过决策树算法得到结果：股票预期收益以三种不同情况加权之后算得的预期收益是 5%，比固定存款利率 7% 低，于是得出投资结论。

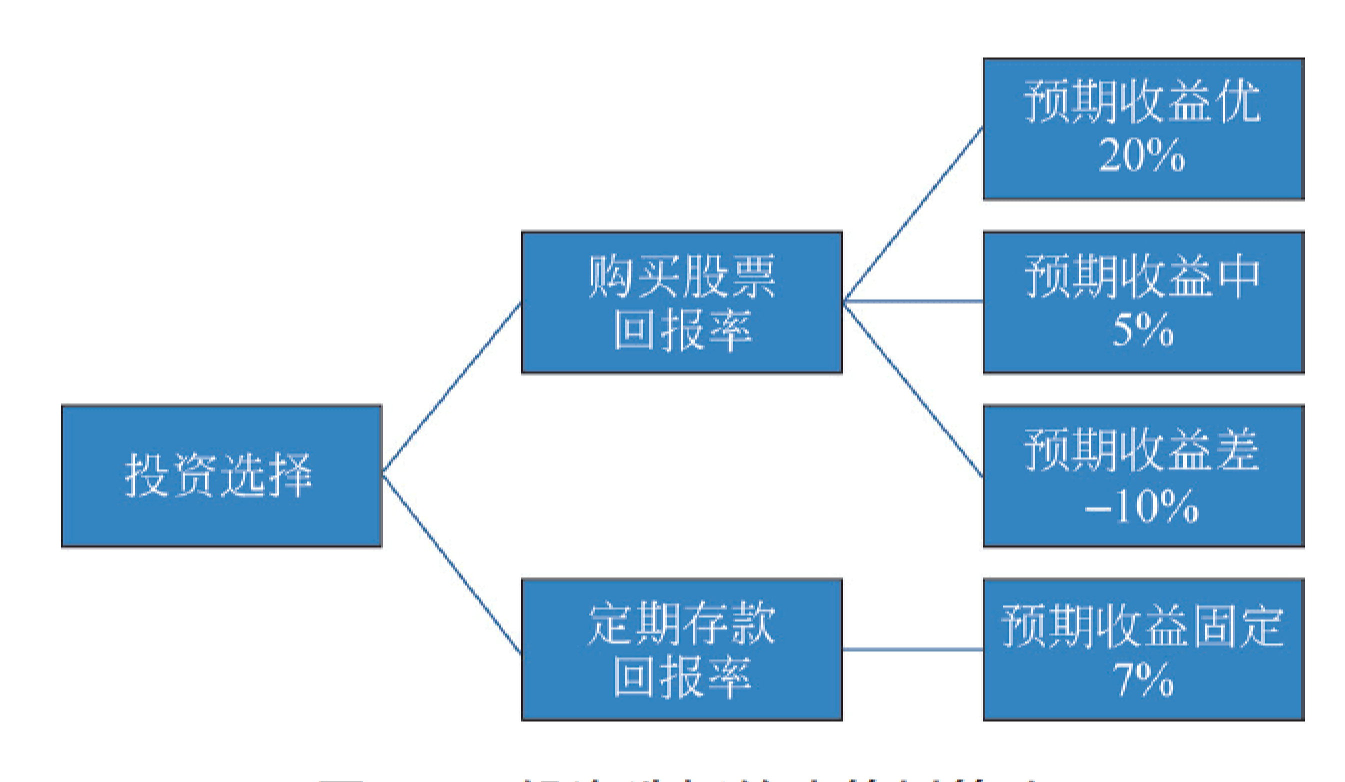


图 3-4　投资选择的决策树算法

以上介绍了数据挖掘中最常见的四种方法，这些方法为处理数据作出基础的分类或者运算。实际的数据挖掘过程并不如这些简化的图看上去这么简单，需要很多计算机科学的知识，但是基本思路的方法是相同的。另外还有很多其他的挖掘方法，限于篇幅在此不作介绍。

3.3.2　数据获取

数据挖掘始于数据吗？错！正如同现代市场营销观念中，销售始于发现消费者需求一样，数据抓取取决于你想要解决什么问题，这就回到上一节中讲到的，弄清楚你要解决的问题，才能知道你需要什么样的数据。当这一步骤很好地完成之后，数据挖掘人员才开始寻找并发现什么样的数据是他需要的，简单地说，能否正确地选择方向和路径，决定了你是否能够到达终点。Pyle（1999）提到，数据获取包含了数据发现、特征描述和数据集成三个阶段。

例如，淘宝网的一个卖家想知道自己店铺的首页究竟应该放多少件商品最好。于是该商家向淘宝网内部工作人员提出这个问题，并愿意为此支付报酬。淘宝网的数据工作人员得到了这个任务，他会首先研究提出该要求的卖家所属类型和业务发展所处的阶段，例如，该卖家是以经营什么商品为主，该商品种类是否繁多而需要单独陈列等。在弄清楚卖家的实际需要之后，数据挖掘人员开始描述他需要的数据具有的特征，而后设定限制条件，这样才能在整个网站数据库中抓取该卖家处于同一商品类别中的其他卖家的页面陈列数据，并且根据一定的关系分为不同的数据集合。数据的抓取实际上建立在对于问题的理解上，只有真正理解问题之后，你所需要的数据类型和特点才能明晰，这样就避免了数据库中掺杂与最终建立起来的商品陈列及购买率无关的数据，从而避免了错误的结论。

另外，数据获取的工具也会影响数据库中数据的形式，从而影响分析过程。例如，遍布在商场各处的摄像头搜集到的是连续的图像信息，有些人称之为数据流，而收银台收到的是一个一个消费时间、内容、金额等数字化信息，为了在后续的数据分析中，使数据形式尽量简单，就需要在一开始搜集数据的终端上改进。

3.3.3　数据存储

数据的存储含义并不难理解，它是把数据流在加工过程中产生的临时文件或加工过程中需要查找的信息、数据以某种格式记录在计算机内部或外部存储介质上。数据存储要命名，这种命名要反映信息特征的组成含义。数据流反映

了系统中流动的数据，表现出动态数据的特征；数据存储反映系统中静止的数据，表现出静态数据的特征。

目前在互联网上每秒钟都产生着大量数据，这些数据以流动的形式存在着。例如一个人浏览网页的过程，鼠标停留的位置，在每个页面停留的时间，从一个页面转向下一个页面的过程都会被服务器记录下来存到一个合适的地方，以供需要研究网民浏览行为的公司或者个人使用。根据信息产业资讯公司 IDC 在 2011 年 6 月发布的“数字宇宙”（Digital universe）报告《从混沌中提取价值》（Gantz et al., 2011），2011 年全球被创建和被复制的数据总量为 1.8ZB。1.8ZB 是什么概念？举例来说，1.8ZB 相当于全球每个人每天都去做 2.15 亿次高分辨率的核磁共振检查所产生的数据总量。[①]

报告还指出，目前这些数据中 75% 是个人制造的，但是相信在未来，全球数据的 80% 将由企业数据构成。这些大量的企业数据则来源于它们对网民行为、网络文本、视频和音频资料的对此分析（Gantz et al., 2011）。这些数据是“关于数据的数据”(Data about data)，计算机科学称之为元数据（meta data）。举例来说，图书馆中各种图书的索书号就是元数据，它可以帮助你不必通过查看图书本身来找到图书。这对大数据时代海量的图片、文字和视频、音频的处理非常重要，用传播学的视角来看，元数据相当于把所有内容“解码”成方便数据仓库存储和查找的单元，进而方便大数据的分析和模型化。比如说在社交网站 Facebook 中的人脸识别可以帮助用户使用“圈人”功能，识别照片中人脸的过程实际上是把图片数据转化为关于人类面部图像的元数据。元数据的增长速度在目前是整个数据量的增长速度的两倍（Gantz et al., 2011），随着技术的发展，新的搜索、发现、分析工具使得元数据产生的速度大大提高。

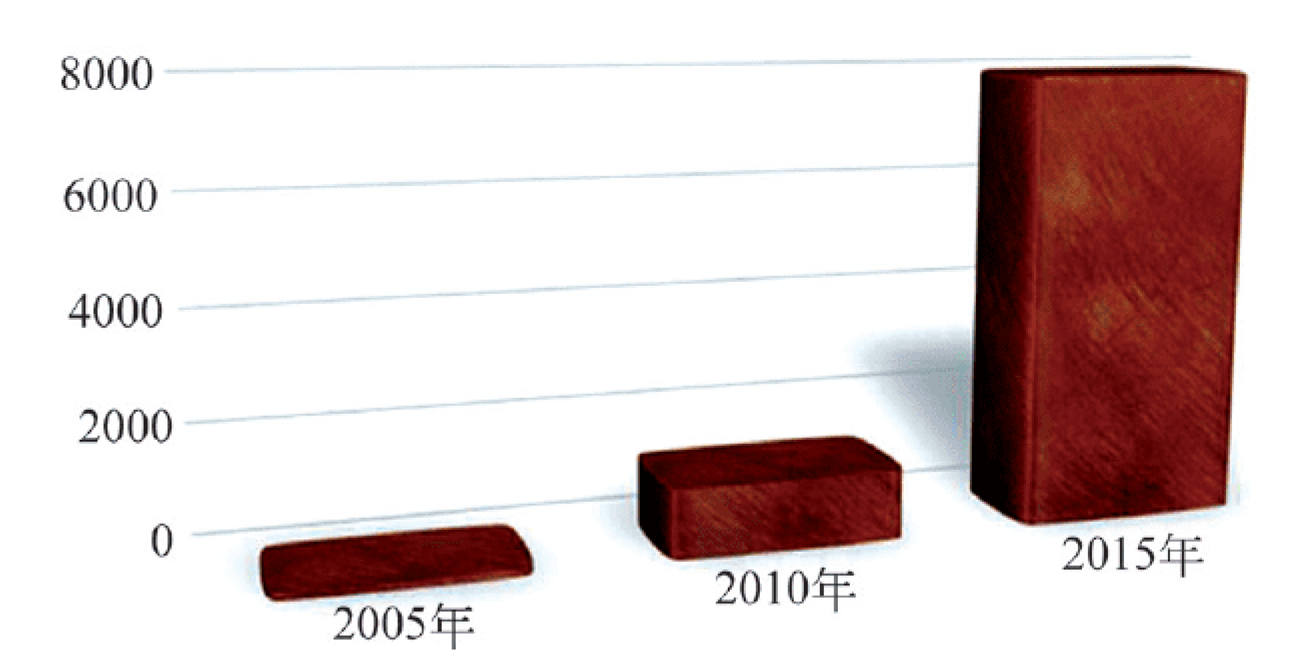


图 3-5　数据存储十年发展（单位：艾字节）[②]

① 这个例子的原文链接：http://storage.chinabyte.com/163/12110163.shtml。

② 资料来源：IDC 报告《从混沌中提取价值》。

爆炸的数据量预示着背后大量的人力和财力的支持。据统计，自 2005 年以来，企业在数据存储上的投资增长了 50%，达到现在的约 4 万亿美元。而存储的平均费用却在降低，未来 5~10 年，相关的人才需求也在成倍增加。图 3-6 显示了 2005—2015 年数据存储平均费用和投资的消长关系。10 年间，每 GB 的存储费用降低到原来的不到 1/10，而总体投资增长不过增加了一倍左右。这个现象印证了摩尔定律（Moore's Law），英特尔公司创始人之一戈登•摩尔于 1965 年在做一份计算机趋势报告时发现，在价格不变的情况下，计算机集成电路上能容量的晶体管数量每 18 个月翻一番。也就是说相同价格的电脑，相隔 18 个月的后者比前者性能翻一倍以上。

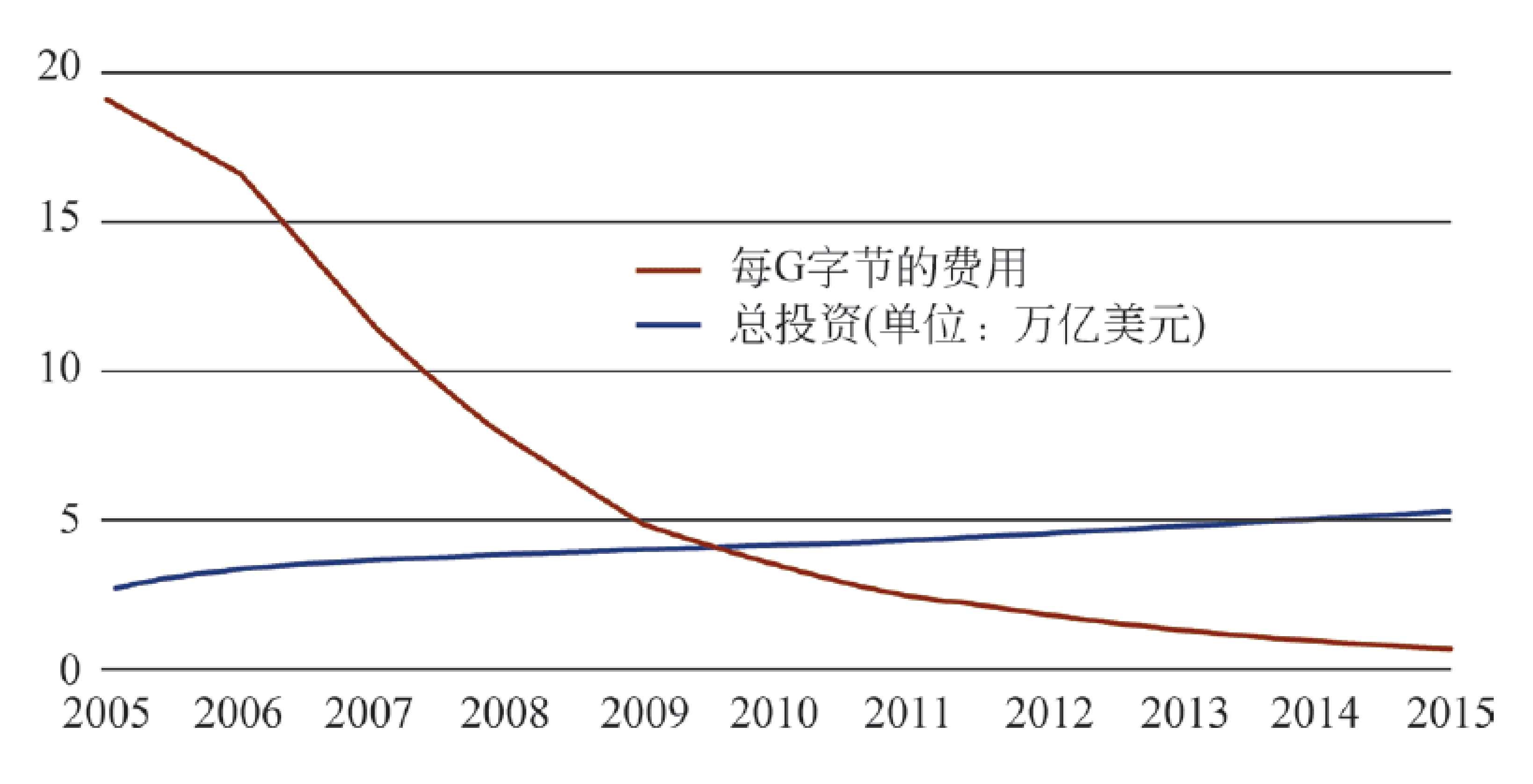


图 3-6 十年数据平均存储费用（GB）和投资（万亿美元）[①]

3.3.4 数据清洗

数据清洗（data cleansing）的概念很容易理解，从字面上看就是把已经存储好的数据中“脏的”数据（dirty data）洗去。更科学的概念是把存储数据中可以识别的错误去除。在数据仓库中和数据挖掘过程中，数据清洗的含义是使得数据在一致性 (Consistency)、正确性 (Correctness)、完整性 (Completeness) 和最小性 (Minimality) 四个指标满足上达到最优，目前数据质量（Data quality）也是在这四个层面定义的。

数据清洗是正式使用数据前最后一道关卡，在数据挖掘领域它也被称为数据的预处理。图 3-7 为数据清洗原理。

① 资料来源：IDC 报告《从混沌中提取价值》。

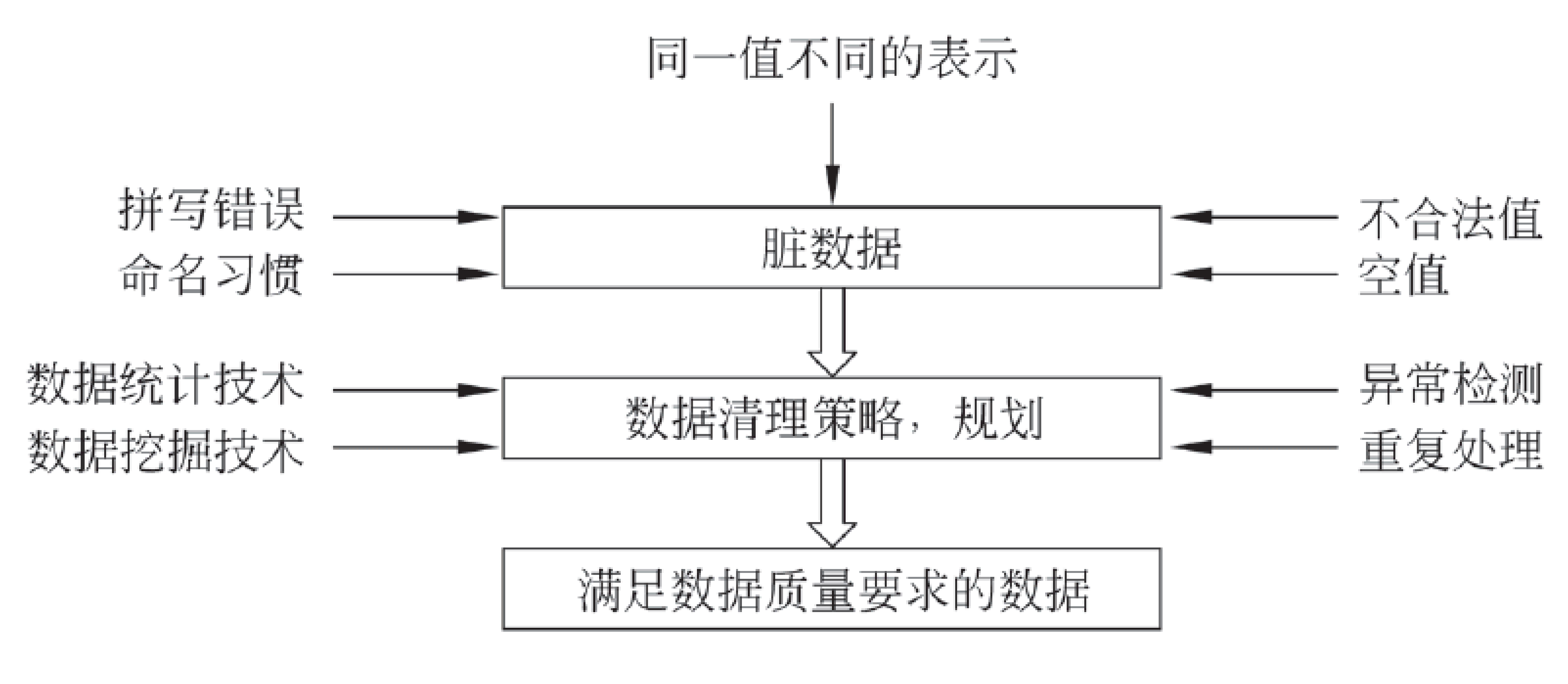


图 3-7　数据清理原理①

大数据背景下，大量来源不一的冗余、复杂、错误数据被存储，之后的“去粗存精”、“去伪存真”工作需要数据清理技术加快发展速度，在极短时间内提高数据质量，满足行业和个人的数据挖掘要求。大数据时代下，人们不缺乏数据，而是缺乏找到有价值数据的能力和工具，这使得数据清洗的价值凸显。不过，目前数据清洗的技术能力还远远不能满足清洗大数据的要求，它或成为数据挖掘技术的一个热点。

3.4　挖掘过程

IDC 的报告《从混沌中挖掘价值》详细描述了数据挖掘的过程。根据存储技术的不同，人们常把数据划分为结构化数据和非结构化数据。简单来说，结构化数据就是能够用统一长度的字段（Field）来表示的数据，如数字和符号。对应的，非结构化数据需要不同长短的字段来表示，这需要数据库的存储和分析能根据需要具有可伸缩性（Scalability）。形象说来，非结构化的数据挖掘技术能同时找出全世界人口的特征分布和对一篇博文的概念与主张的深度分析。结构化数据是过去数据挖掘的主要方向，但是，这些内容只占总体数据量的冰山一角。根据 IDC 的报告，目前所有数据中 90% 是非结构化的数据。这些非结构数据来源于网站上个人发布的文字、社交网络中大量的聊天记录、各种被复制、转发或者重新编辑的 Flash 动画、各种格式的视频和音频等。结构化数据挖掘的一般过程在其他大量商业或者是计算机科学的书籍中能找到，本节重点介绍网络文本挖掘并扼要介绍 WEB 挖掘的过程（Gantz et al., 2011）。

① 杨辅祥，刘云超，段智华等．数据清理综述．计算机应用研究，2002, 19(3): 3~5.

3.4.1 文本挖掘

个人和机构每天使用互联网产生的大量电子文档，比如说一篇有感而发的博文、与好友的聊天记录、转发的微博等。这些文字信息无法使用传统的数据挖掘方法进行分析，文本挖掘（Text mining）是对文本进行的数据挖掘。它最初的应用包括对大量飞机事故报告、警察局档案的挖掘。例如，通过挖掘警察局的案件卷宗，一些地理上分散、时间上相隔甚久的案件之间的联系可以被发掘出来，通过挖掘这些卷宗文本，可以找到零散案件发生的类似之处，或者导致事故发生的共同原因，或者在某个城市的哪些区域和时间案件高发，从而优化警察局巡逻安排、城市管理等。

除此之外，文本挖掘在商业领域也有很大应用。文本挖掘在商业情报（Business Intelligence）应用上得到了很大发展。例如，公司 A 会以自己公司的名字或者某产品的名字为中心搜索所有网络上相关的文本，可能是用户购买之后的评价反馈、博客中体现个人情感的只言片语，还有很多的新闻稿件。文本挖掘包括提炼中心思想、关键词搜索、归纳文章要点、串联各篇文章的主题等；还可以通过文本中语义关键词或者句子搜索信息。在其中，语义网络（Semantic network）是很重要的工具，它通过一系列文本中概念与概念的关系网络来发现最重要的概念。文本挖掘过程实际上是将大量人类语言材料按照计算机语言能够理解的方式分解，再重新组合成具有特定意义的计算机语言然后被人理解，从中发现新的知识或模式。

通常来说，文本挖掘的第一步是找出具有独立意义的信息单元，比如一篇文章中的同义字词。现在已经形成了一个庞大的同义字词库，在此基础上分析文章时产生关联意义，可以帮助人们快速浏览十篇、百篇文章的主要内容。此后建立的文本运算法则将分解的信息重新组合，得出一个总体的模式或者各个关键概念之间的相互关系。目前文本挖掘技术包括自动分类、文本相似性检索（自动排重）、自动摘要 + 主题词标引（自由词 + 行业主题词）、常识校对、相关短语检索、自然语言检索等。

文本挖掘技术在现代信息系统中的应用越来越广泛，其重要性也越来越突出，在信息资源处理的多个阶段，包括信息采集前，后的预处理，信息编辑或加工时的辅助标引、信息服务时的摘要等信息调用参考、信息检索时智能辅助的功能，都需要依赖文本智能挖掘技术来实现。

文本挖掘技术除了在商业领域的应用之外，在今后的公共安全、舆情监测方面的作用巨大。举例来说，关于食品安全的议题近期以来一直是媒体报道和网民讨论的热点。文本挖掘系统可以在所有关于食品安全的新闻报道或者个人

博客、日志中，通过关键词、语义分析网络发现媒体或者是网民群体的态度，究竟是更关心政府的监督还是更集中在批判不法商贩。图 3-8 展示了 30 日内“经济学人”网站中用户评论的关键词及其相互关系。关键词越大说明评论者越多，关键词之间的连线越粗说明同时提到两者的评论数量越多。当用户将鼠标放到某一关键词上时，可以显示与这个关键词最相关的其他关键词和具体的用户评论。这个图为网站浏览者提供了一个直观的印象，了解和他一样的网民都关心哪些话题，并且这些话题之间的关系是什么。当然，这只是文本挖掘的一个小应用。

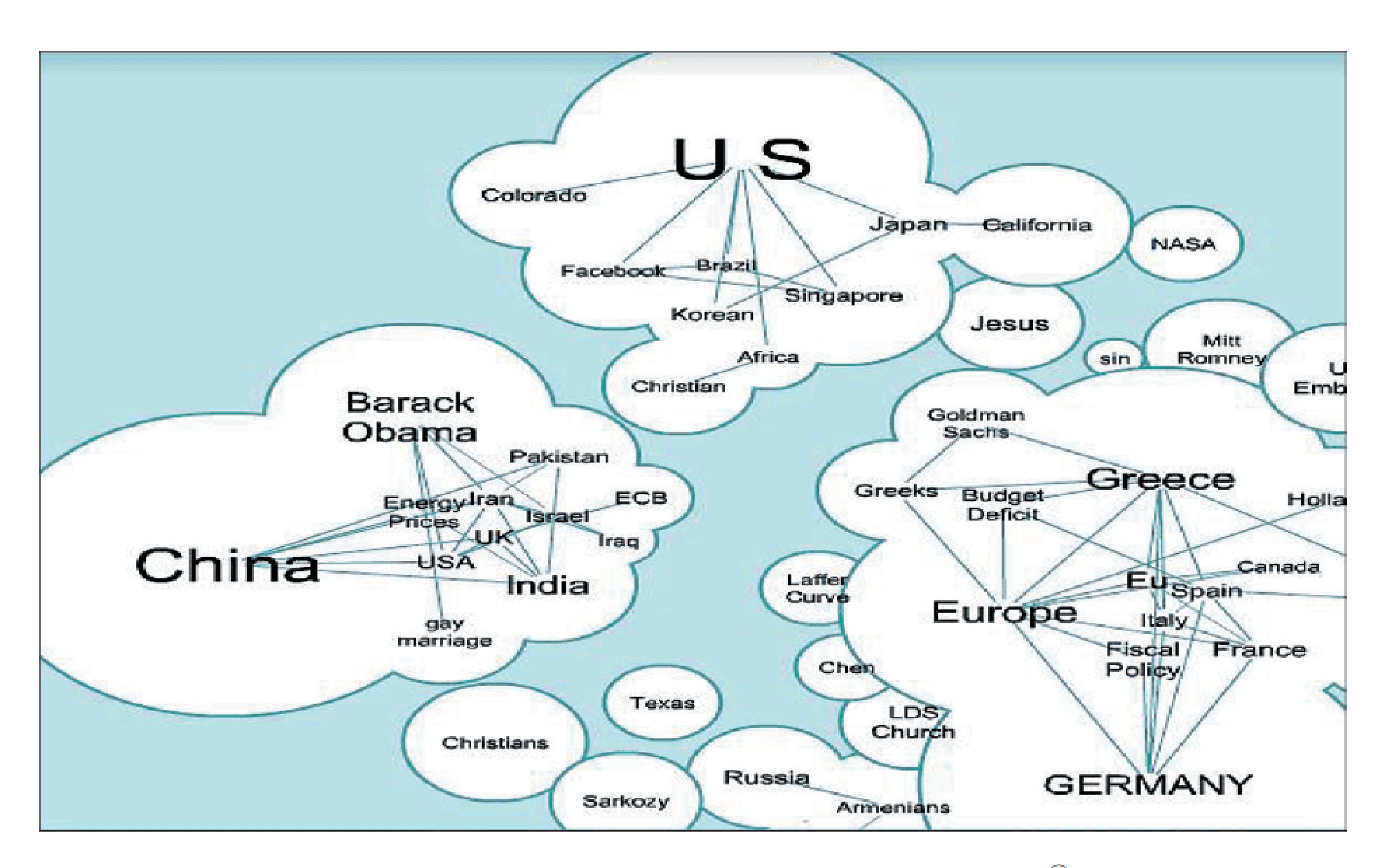


图 3-8　“经济学人”网站评论关键词云截图[①]

上图“经济学人”的“评论关键词云”的例子可以看作是一个小范围的简易舆情检测系统，当一国政府或者公共机构需要实时监测网络热点话题，并对此进行评价和预测是否会发展危机，进而评估这种风险时，就需要用到综合文本挖掘以外其他形式数据的能力，WEB 挖掘就是一个相对文本挖掘内容更丰富、更复杂的例子，它也是未来需要大规模发展的技术热点。

3.4.2　WEB挖掘

比文本挖掘更复杂和更广泛的是 WEB 挖掘。互联网的“信息很密，价值很稀”，人们被大量的信息流所淹没而仍然渴望从中获取知识和价值。WEB 挖掘能够在网络上帮助文件和服务定位。搜索引擎就是这种作用最基本的体现，

① 资料来源：http://www.economist.com/conversation-cloud?days=30，未翻译。

同时也是引导更复杂的 WEB 挖掘形式必要的最初行为，它还包括信息的提取功能，这里的信息是指在搜索行为中的数字或文本数据。被搜集到的数据包括数字、文字、图片以及其他数字形式的媒介。WEB 挖掘在网站上的另外一个关键行为是查看用户行为，研究网络用户的模式不仅有利于预测用户行为，还可以通过一些研究得到的结果改善网站的设计，以达到提高浏览量或者销售业绩的目的。WEB 挖掘的研究分类包含以下三类：WEB 内容挖掘 (Web content mining)、WEB 结构挖掘 (Web structure mining) 和 WEB 使用模式挖掘 (Web usage mining)。

WEB 内容挖掘从网上搜集有用的信息（包括网民访问信息），是指挖掘网页本身所含的内容和网页后台服务器搜集到的网民浏览网页时所留下的痕迹。比如说，购物网站亚马逊（Amazon.cn）首页上会出现“与您浏览过的商品相关的推荐”或者“根据浏览记录为您推荐”的商品。实现这个功能的过程其实是在拥有庞大的用户浏览数据库基础上，了解顾客的购买目标并展开聚类分析，从而预测出顾客还可能想要什么商品。基于 WEB 内容挖掘的好处是数据的真实性和大样本量提高了其结果的有效性。

WEB 结构挖掘是探寻网页与网页之间的关系，或者超链接之间的结构关系。一旦明确了这些结构，数据挖掘人员可以在不同的网站之间方便地查找同类或者近似的内容，或者找到一些更优化的网站设计方式。

WEB 使用模式挖掘是通过分析来自网络服务器的二手数据（元数据）得到关于网民使用网络的路径或者习惯，这些网络使用模式可以是广义的、普通的，也可以根据客户的要求集中地挖掘某一类用户或者某一类网站的使用模式。网络服务日志（Web server log）是取得此类信息的最主要方法，这些信息经常储存在网络数据仓库中，等待进一步的数据挖掘。上文提到的 KDD Cup 的竞赛任务之一就属于使用模式挖掘。通过找到用户点击网页的模式来提高网页投放广告的精准度，是许多互联网企业最关心的话题。

3.5 未来的挑战

人类因为数字才感觉世界更加可靠，并且依赖数字工具探索宇宙的奥秘。在现代，金融、电信等行业业务本身就是对数据的存储、挖掘、传输，或者说靠挖掘数据做出决策。在大数据的背景下，传统的数据挖掘已经不能满足分析和挖掘海量信息的需要了。目前大多数的思路是，以“解码”的方式将大量非结构化的数据转为结构化数据，将多维度的信息以计算机的二维形式呈现，然后再归于结构化的数据处理方式。其中的危险是，“解码”的工具决定了解码

的结果，从而导致一些不能被结构化的信息流失。因此，在未来几年中，数据挖掘的需求可能导致数据搜集、存储方式的改变，这或许会成为信息行业的一个大的变革。

传统数据统计的危险在于，把所有的数据在量上进行比较并得出结论，而它的假设前提是所有的数据都是同质的。这是一个可疑的命题。一些技术人员已经意识到了这个问题，正在开发更加类似于人类自然语言的数据挖掘方式。“异质”数据的挖掘技术才刚刚进入开发阶段。

这只是数据挖掘的一个潜在的危险。而另外一个更迫切、更显而易见的危险是“错进，错出”（Garbage In, Garbage Out, GIGO）。这种危险在于人类现在不缺少信息，或者说是手中的信息太多而无所适从。人类目前抓取和存储信息的能力大增，但是如何辨别数据的价值从而防止大量的错误信息进入待挖掘的数据中，还需要技术的进一步发展。

目前已经有一些公司试图在数据挖掘中挖掘到更多的价值，它们开发的数据挖掘工具开始嵌入到各个需要数据分析的企业的核心，例如 Hadoop 平台、SQL Server 等都开始深入海量数据的挖掘。国内的购物网站之一淘宝网也利用其掌握的一手用户数据推出“淘宝魔方”服务，通过后台数据挖掘用户评论、浏览量、收藏量等来预测某个商家或某件商品的销售趋势。越来越多的个体商家开始求助于销售数据挖掘来提高未来的业绩。

大数据时代的数据挖掘方式必将更加人性化、社会化，以人为中心来改进计算机和互联网技术。这需要改变过去已经建立起来的一些数据存储和传输的方式，如社交网站用户之间的交往模式、上亿张图片被浏览的记录等。大的变革预示着行业未来相关的人才将会紧缺，资金和项目都会大量涌入大数据挖掘业务中。这种业务不是依靠单个公司或者单个行业就足够的，正如人际关系网的交叠，大数据时代下，信息产业和互联网通过大数据挖掘出来的商业价值将制造业、服务业、农业等产业更加紧密地整合在一起。如 IDC 报告（Gantz et al., 2011）设想的那样，大数据时代的技术飞跃需要一个新的“曼哈顿计划”整合公司间、行业间的资源和人才优势。对于中国来说，庞大的用户基数和持续稳定的经济、行业发展状况为大数据挖掘行业提供了优良的发展基础，就像 20 年前，很多人还不知网络为何物、有何用途，而现在没有一天不靠互联网工作、生活一样，下一个 10 年，大数据挖掘、云计算等或许将改变商业的运作模式和人们的日常行为。

这一趋势已经初现端倪。

（本章编译者：汪震，清华大学国际传播研究中心助理研究员，硕士研究生）



第4章 大数据前瞻

大数据的号角已经吹响，它将不可逆转地对人类社会带来革命性的影响，但问题是：大数据的影响将把人类社会引向何方？

“想象一下到2020年我们会是什么样？”美国皮尤研究中心（Pew Research Center）与伊隆大学于2012年7月20日发布了一项调查报告，显示出了人们对这一问题的不同态度和看法。一部分人描绘出了大数据将要创造的乐观图景，能够促进社会、政治和经济的智能化发展；另一部分人则认为到2020年大数据将引发更多的问题。

53%的人对大数据创造的未来世界表示乐观，他们认为：

到2020年，对大量数据集的人为和机器分析将要极大地促进社会、政治和经济智能化发展。大数据的兴起将带来一连串的福音：促进“即时预报”；推动能够评估数据类型的“推理软件”的发展；创造可以从全新角度理解世界的高级算法。总之，大数据的崛起会为社会的各个方面带来积极的影响。

39%的人的观点却恰好相反，他们认为：

到2020年，对大量数据集的人为和机器分析将会引发更多的问题。对于大数据集的分析将会造成我们对自身预测能力的盲目自信，进而会导致很多错误的决定。而且，这些分析结果将会被强权的群体和机构私用、滥用。大数据为所谓的大多数人服务，削弱了少数人的利益，这将会造成恶劣的影响。总之，大数据的崛起对于整个社会而言，无疑是一个噩梦。

无论人们对于以互联网和人工智能为代表的信息革命期待也罢，恐惧也罢，它就那么到来了。大数据的应用已经大大地改变了人类世界，从全球的视野来看，大数据的未来是什么？人们对于大数据的态度有哪些共识与分歧？对于企业家和政策制定者来说，应该如何面对大数据？

4.1 “智慧地球”

4.1.1 “智慧地球”正式提出

2008 年 11 月 6 日，IBM 总裁兼首席执行官彭明盛（Sam Palmisano）在纽约召开的外国关系理事会上，首次正式提出了“智慧地球”（Smart Planet）的概念。



图 4-1 IBM提出的“智慧地球”①

2009 年 1 月 28 日，奥巴马就任美国总统后，与美国工商业领袖举行了一次“圆桌会议”。作为仅有的两名代表之一，IBM 首席执行官彭明盛再次提出

① 图片来源：IBM 官方网站，网址：http://www.ibm.com/smarterplanet/cn/zh/index.html?crs=apch_ccs1_20120229_1330515863559&cm=k&cr=Google&ct=201AG27W&ck=ibm+smart+planet&cmp=201AG。

“智慧地球”这一概念，建议新政府投资新一代的智慧型基础设施，阐明其短期效益和长期效益。奥巴马对“智慧地球”给予了积极的回应：“经济刺激资金将会投入到宽带网络等新兴技术中去，毫无疑问，这就是美国在 21 世纪保持和夺回竞争优势的方式。”（张永民，2010）“智慧地球”被认为是挽救危机、振兴经济、确立美国在未来竞争优势的关键所在，并将上升为美国的国家战略。奥巴马政府在出台的“7870 亿美元经济刺激计划”中，针对宽带基础设施建设、医疗电子档案、电网以及学校 IT 基础设施等“智慧地球”的相关内容制定了战略规划和对其进行了大规模投资（工业和信息化软件与集成电路促进中心，2009）。

彭明盛（2008）在《智慧地球，下一代领导人议程》中指出：这个世界已经紧密相连，不论是在经济、技术还是在社会层面。但是我们也知道仅仅连接是远远不够的。是的，世界会变得越来越“平”。地球变得越来越小，人类联系也将更加紧密。但是，有一点变化潜力无穷。那就是，我们的地球变得越来越“智能化”。这不是简单的一个词，它是指将智能技术应用到生活的各个方面，如应用到各系统和程序之中，以便货物能被顺利地研发、制造、购买，人们能享受各种服务，万物（包括人、货币、石油、水电等）可以顺畅流通，人们可以安居乐业。

智能系统已经彻底改变了电网、供应链和水源管理。它们可以确保药物货真价实，确保外汇兑换安全可靠。它们改变了一切，从公司的业务模型到如何促使员工协作并进行创新等都有了变化。

智能架构逐渐成为国家、区域和城市之间竞争的基础。在全球经济一体化这种大背景下，投资和工作不仅仅只是流向可提供成本利益、技能和专门技术的区域。更重要的是它将流向那些能提供智能架构的国家、地区和城市——高效运输系统、现代化的机场，安全的贸易通道，可靠的电网、透明且可信赖的市场还有更高质量的生活。

4.1.2　智慧地球的含义

IBM 的“智慧地球”战略提出，把感应器嵌入和装备到各种物体中并被普遍连接，形成“物联网”；借助这个整合能力超强的网络，对网络内的人员、机器、设备和基础设施进行实时管理和控制，继而使人类能以更加精细和动态的方式管理生产和生活，达到“智慧”状态。其本质是建立在物联网基础上的更加系统和智能的信息服务，或者说，利用互联网实现物物互联并形成海量数据，再借助专家智慧和多样化的服务模式，为政府、企业等提供便捷的个性化服务或系统解决方案。而 IBM 则希望“智慧地球”推动器进一步向高端服务

业企业转型，“互联网 + 物联网”是实现其转型的工具（张旭等，2011）。

IBM 商业价值研究院的报告《智慧地球赢在中国》指出，“智慧地球”的核心是以一种更智慧的方法，利用新一代信息技术来改变政府、公司和人们交流的方式，以便提高交流的明确性、效率、灵活性和响应速度。如今信息基础架构与高度整合的基础设施的完美结合，使得政府、企业和市民可以做出更明智的决策。智慧方法具体来说具有以下三个方面特征：更透彻的感知，更全面互联互通，更深入的智能化（甘绮翠等，2009）。

更透彻的感知，是超越传统传感器、数码相机和 RFID 的更为广泛的概念。具体来说，它利用的是任何可以随时随地感知、测量、捕获和传递信息的设备、系统或流程。通过使用这些新设备，从人的血压到公司财务数据或城市交通状况等任何信息都可以被快速获取并进行分析，便于人们立即采取应对措施和进行长期规划。

更全面互联互通，是指通过各种形式的高速宽带通信网络工具，将个人电子设备、组织和政府信息中收集和储存的分散信息及数据连接起来，进行交互和多方共享，从而更好地对环境和业务状况进行实时监控，从全局的角度分析形势并实时解决问题，使得工作和任务可以通过多方协作得以远程完成，从而彻底地改变了整个世界的运作方式。

更深入的智能化，是指深入分析收集到的数据，以获取更加新颖、系统且全面的洞察来解决特定问题。这要求使用先进技术（如数据挖掘和分析工具、科学模型和功能强大的运算系统）来处理复杂的数据分析、汇总和计算，以便整合和分析海量的跨地域、跨行业和职能部门的数据和信息，并将特定的知识运用到特定行业、特定的场景、特定的解决方案中，以更好地支持决策和行动。

“智慧地球”的愿景是将世界运行到一个更高的智慧水平，使个人、企业、组织、政府、自然系统和人造系统的交互方式更具智慧。每次交互就意味着有机会以更完美、更高效和更多产的方式完成事件。更重要的是，地球将变得越来越智慧，为人类开创更为广阔的前景（彭明盛，2008）。

4.1.3 智慧地球，从智慧城市开始

IBM 在《智慧城市白皮书》报告中指出，21 世纪的“智慧城市”，能够充分运用信息和通讯技术手段感测、分析、整合城市运行核心系统的各项关键信息，从而对于包括民生、环保、公共安全、城市服务、工商业活动在内的各种需求做出智能的响应，为人类创造更美好的城市生活。

由中国电信智慧城市研究组编著的《智慧城市之路》一书中提到，“智慧

城市”是按照科学的城市发展理念，利用新一代信息技术，在泛在信息全面感知和互联的基础上，实现人、物、城市功能系统之间无缝连接与协调联动的智能自感知、自适应、自优化，从而对民生、环保、公共安全、城市功能、商务活动等多种城市需求作出智能的响应，形成具备可持续内生动力的安全、便捷、高效、绿色的城市生态。智慧城市实现的是城市系统的优化升级，使城市各系统更完善、更智能、更协调、更发达，使城市中的人和物更智慧、更和谐，使城市中的人生活更幸福。

“智慧城市”是一个不断演进的发展主题，是信息技术发展到一定阶段的产物，并随着技术、经济和社会的发展不断持续完善。从技术的狭义角度来看，智慧城市构建了未来城市的技术信息基础，有力地支撑了城市的发展。“智慧城市”带来的改变不仅限于理念范畴，它将对城市的生产方式、生活方式、交换方式、公共服务、政府决策、市政管理、社会民生等方面产生巨大而深远的变革（中国电信智慧城市研究组，2011）。

IBM 在《智慧城市白皮书》中为“智慧城市”概括出以下定义：

“智慧城市”策略就是：在城市发展过程中，在其管辖的环境、公用事业、城市服务、公民和本地产业发展中充分利用信息通信技术（ICT），智慧地感知、分析、集成和应对地方政府在行使经济调节、市场监管、社会管理和公共服务政府职能的过程中的相关活动与需求，创造一个更好的生活、工作、休息和娱乐环境，为了抓住机遇和构建可持续的繁荣，城市需要变得更加“智慧”。

在操作层面上，城市由关系到城市主要功能的不同类型的网络、基础设施和环境的六个核心系统组成：组织（人）、业务 / 政务、交通、通讯、水和能源。城市的组织系统包括公共安全、健康和教育，这些是能否给市民提供一个高质量的生活的重心：城市的业务 / 政务系统代表着业务所面临的政策和管制环境；城市通过交通系统提供给组织和业务 / 政务相互移动的能力；并通过通讯系统来共享信息和沟通；城市也将为经济和社会活动提供两个必要的公用设施——水和能源等。

这些系统不是零散的，而是以一种协作的方式相互衔接，有效地促进执行力和高效性。这六个核心实际上变成了“系统中的系统”。

尽管如此，当重要和积极的转变需要提供潜能时，“系统中的系统”的每个元素都面临着重要的挑战和威胁。例如，城市面临着极其重大的健康保健问题，像婴儿的死亡率、世界各地流行艾滋病问题；对于政务来说，必须在城市系统调整和以满足减少行政费用支出的要求之间找到一种平衡；低效率的交通系统导致运营费用的增加；随着居民和商业通信需求额的增加，城市又面临着挑战；水资源短缺影响社会稳定和生活质量；当前的能源管理监控系统常常不

能提供稳定的检测并且管控效能低下，在安全和效率方面都需要改进。

当城市面临这些实质性的挑战时，当前的模式显然不再是可行的方式。城市必须使用新的措施和能力使城市管理变得更加智能。城市必须使用新的科技去改善它们的核心系统，从而最大限度地优化和利用有限的资源。

智慧城市是一种看待城市的新角度，是一种发展城市的新思路。它要求城市的管理者和运营者把城市本身看成是一个生命体，要求人们认识到，城市本身不是若干功能的简单叠加。城市是一个系统，城市中的人、交通、能源、商业、通信这些过去被分别考虑、分别建设的领域，实际上是普遍联系、相互促进、彼此影响的整体（陈柳钦，2011）。

智慧城市包括四个要素：全面物联、充分整合、激励创新和协同运作。

全面物联：智能传感设备将城市公共设施物联成网，对城市运行的核心系统实时感测。

充分整合：物联网与互联网系统完全连接和融合，将数据整合为城市核心系统的运行全图，提供智慧的基础设施。

激励创新：鼓励政府、企业和个人在智慧的基础设施上进行科技和业务的创新应用，为城市提供源源不断的发展动力。

协同运作：基于智慧的基础设施，城市里的各个关键系统和参与者进行和谐高效的协作，达成城市运行的最佳状态（秦洪花等，2010）。

4.1.4 智慧城市模式比较

日本和韩国的智慧城市基于u（ubiquitous）-city，即以任何时间、任何地点、任何电子装置等可以随时获得信息和服务的环境整体为发展目标。近年来，日本的智能城市也更关注对环境的保护。2010年8月横滨市、丰田市、京都市和北九州市四个地区公布了智能城市的总体规划，其核心是建设生态智能城市。横滨市主要通过大量引入可再生能源与电动汽车，对家庭、建筑物和社区实施智能能源管理。京都市将对各类能源管理的高端技术进行实验。

欧洲的智慧城市则更关注城市的生态环境和智能经济的形成。2009年10月，欧盟公布了新的能源研究投资方案，其中将为智慧城市项目投资110亿欧元，在25~30个城市中发展低碳住宅和交通。一个名为“European Smart Cities”的学术项目认为智慧城市的六个要件是：智慧经济（创意产业）、人才、智慧政府、智慧交通和基础设施、智慧环境和智慧生活。

高科技产业的发展是新加坡、马来西亚等一些东南亚国家建设智慧城市的重点。新加坡提出到2015年打造“智慧国”的战略，其主要内容是建立无处

不在的信息网络，同时发展通信产业。马来西亚几乎以美国硅谷为蓝本建设多媒体走廊，覆盖 750 平方公里的狭长区域，其中将建设 12 个智慧城市。

由于缺乏全国性的政策和标准，中国的智慧城市概念模糊，现阶段更多的是强调 IT 基础设施建设，缺乏整合城市功能的统一平台的建设。智慧城市概念与生态城市、低碳城市、数字城市等重叠。中国的智慧城市多是单一或少数几个城市功能或行业的信息化，以及信息相互连通，而统一智慧平台的建设较少。同时，由于缺乏智慧城市所需的产品和技术，许多城市将建立智慧城市相关产业作为重点，同时也借此吸引新投资。为解决中国高速、大规模城市化带来的各类问题，针对城市生态改善和公共服务的应用是中国智慧城市建设的核心（中国三星经济研究院，2011）。

在新经济形式下的城市化变革中，对于城市规划和管理、社会稳定与社会安全、就业、民生、可持续发展等各方面都提出了严峻挑战。致力于支持政府建造更加智慧的城市，公民和企业可以便捷地享受优质的公共服务，各种突发事件可以被迅速有效地应对、解决，保证城市各个系统高效顺畅地运转以及可持续地发展。

4.2　公众意见的分歧

美国皮尤研究中心与伊隆大学关于“大数据对未来的影响”的民意调查显示出了人们对大数据态度的分歧，53% 的人持乐观态度，39% 的人却认为大数据的崛起对人类而言是噩梦。

以下是受访者主要观点的集纳（Janna et al., 2012）：

4.2.1　积极态度

一种主流的乐观态度认为，到 2020 年，大数据的运用将增进我们对于自己和世界的理解。“媒体和监管者妖魔化了大数据及其对隐私的威胁，” Jeff Jarvis 教授说，“这些道德恐慌通常是由技术变革引发的。但是还存在这样的道德事实：我们能够在大数据中发现价值，并且在最新发现的公共性中找到价值。谷歌的创始人已经呼吁政府监管者不要要求他们快速删除相关搜索，因为他们已经能够在医疗官员之前追踪到流感疫情。而且他们相信类似的传染病疫情追踪将拯救数百万人的生命。妖魔化数据，不论是大数据还是小数据，就等于妖魔化知识。这显然是不明智的。”

Mead&Clark 分析师 Sean Mead 认为：“大规模的公共数据集、更便利的工具、分析技术更广泛的推广、初阶段的人工智能软件将点燃经济活动的爆点，

相比 20 世纪 90 年代的互联网和个人电脑革命，能更大地提高生产力。”

哈佛大学柏克曼中心 David Weinberger 认为：“我们刚开始对大数据可以解决的问题有所了解。我们掌握的知识将会囊括我们之前所不能理解的更多真理，因为我们人类的大脑是如此狭小。”

企业家 Bryan Trogdon 说：“大数据就是新的石油。对于企业、政府和机构来说，资源的开发意味着极大的优势。速度、敏捷和创新决定着谁输谁赢，大数据使我们从过去的‘劈一次柴前测量两次’的思维模式转向‘快速下小注’。”

同时，乐观态度还相信“即时预报”、实时数据分析和模式识别必定会更加完善。

谷歌首席经济学家 Hal Varian 认为：“我是即时预报的忠实支持者。几乎所有大公司都有实时数据库，掌握有比政府机构更多的即时经济数据。在未来的十年，政府也可以利用私企的数据。这将会推动制定更加全面、主动、有效的财政和货币政策。”

迈阿密大学环境工程学生态经济主任 Gina Maranto 认为：“全球气候变化的情况下，为了使我们的社会对人为污染和极端天气条件更加敏感和适应，即时预报势在必行。然而，光有数据还不行，我们必须对决策制定有更加深刻的理解，这需要我们扩展对意识偏见及多方合作的理解。”

电影制作人 Tiffany Shian 认为：“大数据使我们看到之前从未看到的模式。这种相互依赖和联系会带给我们一种全新的看待问题的方式。它使我们能实时看到我们行为的结果和影响。我们买什么、吃什么、扔掉什么都会呈现在实时地图中。我们可以及时看到自己行为造成的连锁反应。”

乐观态度还认为，尽管大数据的应用必然会带来一些负面影响，但总的来说是利大于弊。“互联网放大了日常生活中的好、坏和丑恶”，微软高级研究员 Danah Boyd 说，“当然这些会被善意或是不怀好意地利用。科幻小说为我们提供了无数想象的模板。但是二分法却不会给我们带来任何好处。有意思的是，经济交流和信息共享如何发生变化，会为我们开启未曾想象过的可能性。这意味着我们将失去已有的一部分，迎来新的可能性。”

4.2.2　对大数据的忧思

持相反观点的人们对大数据的未来抱有悲观态度。

通讯专家 Oscar Gandy 说：“有必要多思考一下损害在大、中、小型数据收集者、代理者和用户之间的传播。如果大数据的运用是处于社会效益而非利益追逐，那么我将支持数据未来及物联网。”

Wolters Kluwer 的高级分析师 Marcia Richards Suelzer 说："我们可以在纳秒之间做出灾难性的错误计算，并将其推广开来。"

受访者担心政府和企业没有分享信息的动力。对他们来说，监管才是大数据的核心。

GlobalSecurity.org 负责人 John Pike 说："这个世界太复杂了，很难嵌入到如此无差别的大数据概念中。我们讨论的到底是谁的大数据？华尔街的？谷歌的？还是国家安全局的？我本身无比渺小，对于大的概念不感兴趣。"

另外一个匿名人士说："数据整合在今天只有两个原因：国家安全装置和市场数据库。这都不是从网络个体用户的利益出发，反而要么将这些用户当做潜在的恐怖主义分子，要么将他们当做产品和服务的消费者。"

另一位匿名受访者说，"金钱是大数据发展的主要刺激因素。最后的结果很有可能是，大数据将聚焦在使目标群体消费更多的产品。在我看来这无益于社会的发展。我不会说这是一种滥用，但是这是一种利己主义。"

还有一种担忧是，富人将会从大数据中获益，穷人则不能。

加州大学伯克利分校讲师 Brian Harvey 说："大数据是在牺牲穷人的基础上，使富人获益。我认为有小部分人会把这看成是积极因素。你们给出的两个选项'造福社会'或者'危害社会'应该改为'富人获益'或者'穷人获益'。根本没有什么社会可言，只有富裕、贫困以及阶级斗争。是的，我知道非洲的农民用手机追踪大城市里的产品价格。这是好的，但还不够。"

一些受访者担忧大数据会被滥用。

一位校长 Heywood Sloane 说："这并不是互联网或大数据的问题，而是什么人、多少人将会滥用大数据的问题，不论有意还是无意。这样的问题一直存在，所以需要能抵抗滥用的力量：竞争、透明、监察，等等。当然有时也会判断错误。"

教育家 Tom Rule 说："永远不要低估人性的愚蠢和罪恶。"

一位匿名受访者说："数据被滥用有很多原因，解决方法不是去控制数据的收集，而是加强大家对数据滥用的教育，提高合理使用数据的意识。"

4.2.3　喜忧参半的看法

来自纽约的研究咨询媒体公司的负责人 Stowe Boyd 认为："总体而言，物联网和大数据推动了我们感知、理解和控制这个世界的能力。但是，潜藏的分析机器仍然需要人类识别和管理。就像最明亮的光线也会投下最黑暗的影子一样，大数据也有其黑暗的一面。它为不端的应用创造了机会，比方说监控社会，

官方监视和分析我们的一举一动。另一方面，大数据带来了福音。社会宣传群体可以低成本甚至零成本收集到需要的信息，这在今天是不可能做到的。以避开国际粮食组织控制、将区域内的事物生产者和消费者联系起来的草根发明替代事物网络为例，这种被称为事物科技的系统，很可能是基于人们的消费、农民的生产计划、区域内的物流工具等公开信息之上的。所以像所有人类的科技变革一样，大数据也是让人喜忧参半。”

未来研究院创始人 Jerry Michalski 用一种实际的方式描述了大数据的好与坏：“回想起来，人们总是认为自己比实际知道得多。比如说，我们对于科技效果的理解要滞后于对其实施效果的理解。人们最好的意愿是用大数据解决大问题，但是却不能如愿达成。好的创意已经引发了无数糟糕的决定。想想多米诺骨牌效应、优生学和种族优越论，甚至是优胜劣汰。正是这些理论使我们不断犯错。同时，恶意使用大数据将会产生巨大的危害，从对人口的隐形操控到各种形式的隐私侵犯。那些反乌托邦的电影离我们的现实不再遥远。fMRI 实验中得出的数据使我们确信我们明白人们如何做出决策，从而导致我们做出了更多的错误政策。当然也有希望。当人们开始围绕真实数据一起工作时，他们会取得真正的进步。Wikipedia、OpenStreetMap、CureTogether 等都是互联网时代的产物。我们需要大数据创造更多的小群体，与各地的小群体一起合作，创造更加实用的产品和服务。谷歌已经运用大数据找到了解决拼写检查和翻译等棘手问题的简单方法，更不用说疫情追踪等。我畏惧谷歌的巨大威力，但却敬佩他们清晰简洁的方式。”

《未来》杂志副主编 Patrick Tucker 认为这些变革为“可知未来”增加了新维度：脸谱和谷歌之类的服务可以帮助我们更好地理解生活。但是脸谱对于我们生活和社交圈的观点要远远比我们自己的看法更加清晰。问题来了，还有哪些人可以使用这架显微镜？同时，也有很多问题随之而来。Moveon.org 总裁 Eli Pariser 在他的新书《过滤泡沫》（*The Filter Bubble*）中将其描述为“信息决定主义”，是网络过于个人化不可避免的结果。“你过去的点击将决定你未来将要看到的内容，这种互联网历史将不断重演。你将陷入一种停滞的、越发狭小的恶性循环。”

谷歌和脸谱仅仅是最明显的麻烦，因为它们使用那些数据向你提供服务。但是你也可以选择退出脸谱，事实上已经有数百万人不再使用脸谱。虽然将谷歌清除出你的生活并不像十年前那么容易，但是可以匿名使用谷歌，也可以不用谷歌就可以找到信息。这是我们可以选择进入或退出的网络。

未来机器接管了创造未来的任务。他们的预测将对现实世界造成影响，因为个人、群体和国家对于未来的互动成为个人和国家身份的一种表达。不管将

会发生什么，未来作为一种概念，将塑造我们的消费、投票和社会行为。未来越来越易知。我们正站在科技巨大革命面前。

4.3 企业领导者如何迎接大数据时代的到来

麦肯锡的大数据报告指出：随着大数据的价值日益增长，针对大数据的智能开发也成为企业竞争的关键。我们已经看到许多企业已经先于其同行业者在公司绩效评估等过程中运用大数据。大数据将成为不同领域间竞争的重要基础，所以将大数据列入其商业计划势在必行（James et al., 2011）。

不同领域运用大数据创造价值的机会不同，麦肯锡的报告建议，要想充分利用大数据的力量，还要解决如下一系列的问题。

4.3.1 库存数据资产：专利、公开、购买

随着数据变成主要的竞争资产，领导者必须了解自己持有和可以使用的资产。机构应该建立专利数据的存货清单，对有可能使用的数据进行系统编录，包括公共数据（比如政府数据和公共领域发布的数据）和已经购买的数据（比如从数据价值链中的数据收集者处买来的数据）。

确实，要获得转变的机会，企业越来越需要从第三方处获取信息，并将这些信息与自己已经掌握的信息进行整合。在有些情况下，企业还需购买数据的使用权。在另外一些情况下，第三方并没有分享信息的意愿。这时，企业应该充分考虑并拿出极具吸引力的条件，说服第三方将信息出售或共享，或者拿出其他的刺激因素确保数据的获取。要想能够持续、安全、及时地获取外部数据，还需解决一系列的技术难题（比如数据标准化以及数据传输）。

例如，许多企业最近发现了社交媒体的数据价值。电信公司发现社交网络中的信息可以帮助预测顾客走向。它们发现当一些客户开始使用某种通讯方式时，其他客户也会模仿，所以这些容易受影响的用户则被定为保留项目的目标群体。另外一些企业发现它们可以从网上的情绪表达总结出客户的态度、购买趋势和品位，使它们能够及时改变营销策略和产品设计。

4.3.2 明确潜在价值的机遇和挑战

为了抓住机会，进行有目的的实验是充分利用大数据最强有力的途径。选择一些潜力巨大的领域用大数据进行实验，比如数字化营销，然后对成果进行快速分类，将是开始转型的有效方法。

在麦肯锡的研究中，发现创造大量的新价值不一定非要直接进行复杂的大数据分析。在很多情况下，即使是在采用更先进的工具之前，关注数据的使用和应用基本的分析方法就能创造巨大的价值。在医疗保健领域，创造透明度和进行基本的大数据运用就可以产生 40% 的潜在价值。大多数机构正在逐步培养这种能力。

麦肯锡的研究明确了大数据研究和应用的四个阶段。

第一个阶段是数据的数字化与构建，这是使用大数据之前的阶段，是确保产生、构建和组织数据从而使终端用户和后续分析都可直接使用数据的几个步骤。这些技术包括净化数据以排除错误、保证数据质量，将数据结构化，加入描述数据的元数据。

第二个阶段要求数据可以通过网络而被使用，这将成为提升自身价值的强大动力，同时也是数据整合的重要初始阶段。

第三个阶段则是基本分析的应用，涵盖了许多方法，包括基本数据对比和相对标准化的定量分析等。

第四个也是最高阶段是高级分析的应用，比如说自动计算和实施数据分析。此阶段通常可以创造最关键的新的商业模式。它们也允许新的实验手段，针对客户设计最优方案，并且与第三方一起创造更多的新机会。这一阶段需要深层次的专业分析能力。

除了检测机构的潜力，领导者还可以通过通过数据类型检测不同的机会。通过机构已经有的专利数据集抓取的价值，尤其是通过额外的分析，可以带来许多机会，即第一种机会。例如，医疗保健服务的提供者可能发现通过分析临床结果可以更好地确定医疗事故的原因。第二种机会来自于在分析中加入新数据。这些数据通常涉及非标准的数据类型。比如说，保险公司可能发现加入远程感知数据可以更好地评估房地产的风险。第三种机会来自基于大数据建立起来的新的商业模式。例如，支付方发现通过出售基于支付处理过程大数据流产生的客户信息可以创造新的业务。

大数据可能会带来潜在的威胁。在大数据价值链的语境中，信息集合和分析正变得越来越有价值，所以数据生成者要更加充分地理解潜在价值，并且抵御聚焦在数据集合和分析方面的新对手。在一些领域，第三方数据收集者自身并不产生数据。他们在数据聚合服务的基础上提供其他的附加值服务。比如说，在金融领域，网上公司 Mint 聚合个人的金融数据并且提供附加值服务（比如金融计划工具），即使它自身并不产生金融数据。Mint 与其他使用这种商业模式的公司对通过对客户整个金融情况的全面掌握而建立客户关系的传统金融机构构成了威胁。

事实上，对数据等级和 IT 基础设施的需求可能会成为巩固已有成绩的关键动力，在小规模参与者遍布的领域中，机遇和挑战并存。医疗保健领域则是一个很好的例子，许多相对小规模的医疗从业者依然存在。当他们进入到电子医疗病历的数字化时代，并且开始从数据中获取利益时，他们会发现通过与其他从业者合并以扩大规模能获取更大的利润。

云计算推动的数据获取也有可能打破已有的商业模式。许多分布式的共同创造涉及与外部伙伴和客户的合作从而行使更多的公司职能，从研发、市场营销到客户服务，这些传统意义上都是由内部员工完成。

4.3.3 增强内在能力

商业机构需要找到合适的人才,从大数据的应用中获取价值。在人力方面，麦肯锡的研究表明，越来越短缺的人才将是那些能够分析大数据的深层分析人才；知道如何利用大数据分析结构的管理者和分析师；进行大数据操作的支持性技术人才。

有些大数据的公司已经对深层次的分析人才进行了充分的分类，其他机构可以从这些最佳实践中获取很多经验。鉴于对人才的潜在竞争，机构必须大规模地招聘这类人才。这包括从其他公司挖人或者从其他公司购买分析服务。值得注意的一点是，早期招聘的人员非常关键，因为他们构成了团队。让这些人招聘可以替代他们的人很难，所以在早期就招进最有能力的员工是组建一支高效团队的最好方式。

领导者还需要弄清楚如何组织这些深层分析人员，从而使他们形成一个人才中心，能够与其他部门有效沟通，与领导者高效合作。同时，对这一人才库的激励需要金钱，更重要的是内在激励因素。

但是，仅仅拥有人才库并不足以完成机构的转变，尤其是在领导者和分析师不知如何利用大数据的情况下。所有的领导者必须对分析技术有基本的理解，从而有效利用这些分析结果。机构可以将这一因素考虑在内，调整招聘标准。更重要的是，他们需要创造新的培训计划从而提升现有管理和分析水准。例如，基本的数据项目或者在当地大学的一系列数据分析课程可以激励出一支能够引领机构转型的管理者和分析师队伍。金融公司 Capital One 已经成立了内部培训机构，还提供了实验设计等专业项目。机构必须合理分工，整理激励因素、结构和工作流等，从而使各个层次的员工都可以充分利用大数据带来的信息。英国零售商 Tesco 已经形成了一套从高级管理层到生产第一线的数据导向思维模式。它通过各种以客户为导向的大数据策略，将客户信息整合到所有

的操作环节。鞋业零售商 Famous Footware 的行政团队每两周与试验负责人开一次会，讨论结果并计划新的数据收集和评估项目。Amazon.com 开除了一个网站设计组，因为他们没有经过对客户行为的实验调研就对擅自改变了公司的网站。在这些公司中，大数据成为管理层对话以及公司文化不可或缺的一部分。

4.3.4 推进实施数据策略

为了迎接大数据时代，商业机构应该为整个企业制定一套完整的数据策略，从整体上考虑数据模型、构造以及解决方法的属性。以客户数据为例，最普遍的问题是分散的单位可以在不分享或整合机构数据的前提下形成自己的数据策略。结果是，机构通常发现它们甚至对于自己的客户都没有一个清晰的概念。即使在同一个单位之内，也存在这样的差别。缺乏以客户为中心的意识严重制约了机构使用大数据创造新价值的能力。一个有效的企业数据战略必须包括能够构成彼此协作关系的数据模型、数据交互架构、整合架构、分析架构、安全性和遵从性以及一线服务。

许多机构需要对 IT 硬件、软件和服务进行投资，从而可以抓取、存储、组织和分析大规模的数据集。投资水平因公司现有的 IT 能力而异。IT 领导者需要评估技术差距，以有效地捕捉、存储、积累、交流和分析数据。他们需要与公司内的其他领导者一起合作，研究商业案例，进行新的投资。

尽管需要一个全面的企业数据战略，目标项目的实施和能力的发展也是非常有帮助的。不如说，加利福尼亚州的 Kaiser Permanente 最开始通过建立疾病登记和专家组管理方案，专注于专门为长期病患服务的 IT 项目，而不是能够解决一系列问题的全面 IT 方案。

4.3.5 解决数据安全等问题

随着越来越多的数据因各种目的而无障碍地流通，解决隐私和安全问题将成为重中之重。尤其是隐私权，不仅是法律法规，也是机构与客户、合作伙伴、员工以及其他利益相关者之间建立信任关系的基础。公司务必要制定符合隐私法律法规的数据政策。但是，在制定隐私政策的过程中，机构需要充分考虑要与利益相关者建立何种法律协议，而且需要与利益相关者进行清楚的沟通，尤其是客户，因为他们越来越担忧自己的信息将如何被使用。

作为企业数据战略的一部分，企业需要实施涵盖全部 IT 部门的风险战略。这一战略必须包括深入的企业风险评估，从实体闯入的可能性到黑客侵入的可能性，但是也许更加重要的是，有权使用这些数据的人做出违反公司意愿的行

为。有一系列的 IT 方案专门用于解决数据隐私和安全风险的问题。

企业领导者还需要全力对付涉及数据专利权和责任的法律问题。这些问题需要专业的法律顾问，也需要将多种因素考虑在内的途径，包括战略、与客户、合作伙伴和雇员的关系以及技术等。

运用大数据竞争和获取价值需要领导者扫除各种障碍，包括人才、技术、安全隐私、机构文化和获取数据的激励因素。

4.4 对政策制定者的建议

麦肯锡报告指出：没有政府与政策制定者对于当前大数据发展面临的困难和挑战的回应，将无法发挥运用大数据获得的价值潜能。研究表明，大数据不仅是单个企业强有力的竞争手段，而且可以提升整个行业和经济体内的生产力、创造力以及竞争力，不论是在发达国家还是在发展中国家（James et al., 2011）。

具有前瞻性的政策制定者将与大数据的发展步调一致，并及时找出扫除创造价值过程中障碍的方法。如果政策制定者想要帮助企业最大程度地利用大数据，则需要国家和国际层面的措施。运用大数据将使企业在竞争中脱颖而出，大数据也将在国家间的竞争中发挥重要作用。

政策制定者必须选择那些有助于企业通过运用大数据创造价值的措施。政策可以发挥作用的领域包括建立人力资本、保证数据使用的刺激因素、解决隐私安全问题、建立知识产权框架、克服获取数据的技术障碍、促进信息技术基础设施的完善。一些政府的政策制定者已经开始在这些领域里着手解决这些问题。

4.4.1 为大数据时代建立人力资本

政府可以通过很多方式增加大数据的人才供应。首先，政府通过教育杠杆为社会输送大数据所需专业领域的毕业生。例如，在美国，从联邦、州到当地都有支持科学、技术、工程和数学教育的推动政策。但是，对于深层分析人士的需求更加具体，更多接受过统计学等方面教育的毕业生供不应求。政府增加人才供应的第二种方式是减少人才流动障碍，比如通过远程工作或者鼓励人才移民。

为商业、政府和社会机构培养具有基本分析技术的人才，是一个巨大的挑战。新知识青年至少应该接受这些领域的课程；统计学的必修课也应该成为商业管理等其他管理课程的一部分。但是仅仅等待新一批的毕业生是不够的。政

府应该创造更多条件对已有的管理者和分析师进行培训。

麦肯锡研究发现，很难找到美国之外的国家劳动力的具体信息。一些国家应该提供相关职业的具体数据。基于“不了解则无法管理”的原则，这些国家的政策制定者应该促使劳工统计机构开始收集更多、更详尽的具备高级知识的劳动者的就业信息。这些数据能够有利于更好地打造人力资本。

4.4.2 促进数据分享，创造激励因素

运用大数据创造价值的很重要的方式是对多方数据进行整合。但是很多情况下数据市场并未发展起来，或者已存的数据交易市场失灵。政府可以为市场的有效运作创造条件，包括设立有关知识产权的规则、调停争端等。比如，美国医疗保健领域建立医疗信息交换的要求就是为了保证清洁的临床数据可以得到共享，从而使整个行业可以充分利用有关治疗对比效果的数据。

当市场失灵时，比如利益相关者缺乏共享信息的利己因素时，政策制定者应该制定规则保证信息的共享。比方说，很多企业因为害怕对自己的名誉造成损害而不愿公布失误数据。但是政府却有明确的理由促进信息共享，因为这样可以降低行业范围内的风险。下命令收集和公布此类信息是必要的。例如，政府可能要求公共企业提供标准电子模式下的金融数据。在最近的全球金融危机之后，许多政府已经意识到提升金融报道的透明度有利于降低金融系统的风险。

4.4.3 平衡企业创造价值的需求与大众保护隐私的诉求

虽然大数据可以创造巨大价值，但是很多人对于高度私人化信息的使用抱有怀疑态度。大众将持续表达对隐私权的诉求，企业需要清楚地知道什么信息可用，什么信息不可用。在有些情况下，个人信息市场可以发展起来，但是在另外一些情况下，传统市场机制不足以保护隐私。

在未来，《保护隐私法》的颁布和有效实施至关重要。这不仅有利于保护客户隐私，同时可以证明信息分享的价值大于其风险。政策制定者面临的挑战之一是与大数据的发展保持步调一致。当然，政府、非营利机构和私人企业都需要开展相关的教育，使公众明白有多少私人信息可以共享，这些信息用于何处、如何使用，以及个人是否乐意共享私人信息。

大多数发达国家已经有了专门负责制订和推行数据隐私法律法规的机构。在美国，联邦贸易委员会将《公平信息实践原则》作为处理安全隐私问题的指导方针，欧盟有《信息保护条例》。所有这些法律法规都包含了相似的保护条例。

德国有专门对行业信息保护进行监管的联邦官员。韩国的信息保护法案则是由两个不同的政府部门联合实施的。

与此同时，企业和政府都需要强有力的法律防止黑客和其他入侵，最大程度地保护数据库的运转。保护 IT 基础设施意义重大，在网络袭击日益复杂和猖獗的情况下，要确保可以安全使用数据。例如，针对一个国家金融基础设施的网络攻击会造成数百万人敏感的个人信息的泄露，也会使用户对电子市场失去信任。

4.4.4　建立有效的知识产权框架，保护创新

毫无疑问，在大数据时代，我们会继续见证更多创新沿着数据价值链兴起，生产、抓取、分析数据的创新技术将会不断涌现。随着机构对于大规模数据的实时抓取、存储和分析日益强烈的需求，相应的存储和分析技术也会不断提高。这些创新则需要有效的知识产权体系，既可以保证有价值信息的不断产生，也可以对不同的信息进行有效的分享和整合。人们对于保护知识产权以及争端裁定的需求将越来越强烈。

4.4.5　清除技术障碍，加速关键领域的研发

政策制定者可以帮助解决大数据使用的相关技术问题，包括加快 IT 工具的标准化制定，鼓励关键领域的研发。

IT 工具和某些数据类型的标准化至关重要，这关系到能否通过数据共享创造价值。这些标准源自行业标准的制定，但是政府也发挥了重要的推动作用。例如，在美国的医疗卫生领域，由国家医疗信息技术办公室公布的“电子医疗病历标准”明确了电子医疗病历技术的资质标准，从而使医生和医院可以放心应用该系统。

政策制定者也可以加快大数据研究。政府可以直接发起基础性的研究计划。比如，美国国家科技基金会赞助了计算机科学和数学项目；欧盟推出了研究框架项目，专门用于为欧洲范围内的科技项目提供研究资金。

政府还可以考虑如何通过包括税收和其他金融支持在内的激励因素，帮助克服大数据使用过程中的技术障碍。有时针对大数据的投资和回报存在不对称的问题。仍以美国的医疗领域为例，医疗服务提供者是电子病历技术的主要投资者，但是其产生的利益多由患者和支付者享有。2009 年推出的《美国复苏和再投资法案》向医疗服务者提供了 200 亿美元，用于电子病历和医疗信息共享的投资，从而收集更多的临床数据。

4.4.6 确保对信息和通讯技术基础设施的投资

大规模数据集的运用需要到位的基础设施，包括支撑信息技术的电路和数据传输需要的通讯网络。麦肯锡通过对不同国家的研究发现，鼓励基础设施的政策干预存在很大差异。

许多国家已经推出了扩展基础设施的激励计划。比如，美国政府公布了一系列的财政刺激计划，鼓励宽带基础设施的建立和电子病历的发展。美国政府还提出一项影响深远的国家无线发展规划，计划使 98% 的区域可以使用 4G 宽带。其余各国政府也纷纷采取措施促进基础设施的发展。比如，韩国为某些群体的宽带使用提供补贴，而日本和欧洲国家则明确要求宽带用户需有偿使用网络。

政策制定者可以确保企业等机构充分发挥大数据在人才、研发和基础设施等重要领域的潜能，也可以促进这些领域的创新。这些政策应该包括具体的、可行性强的措施。更加复杂的挑战则是确保立法机构在允许自由运用大数据和减轻公众对隐私安全的担忧之间达到平衡。这一平衡需要透彻的思考，是政策制定者逃不掉的问题。

（本章编译者：刘沙沙，清华大学国际传播研究中心助理研究员，硕士研究生）



第5章 大数据安全

由大数据所带来的巨大利益对企业具有无限的吸引力，自互联网诞生之日起，非法收集互联网用户资料、黑客侵入电脑终端等严重威胁数据安全的行为不断发生，在大数据时代，需要制定和遵守一定的“游戏规则”，保护公众隐私与国家的非传统安全。

5.1 公众隐私与信息安全

5.1.1 个人信息与商业机遇

互联网如今已经从仅仅满足于大众化的信息发布，更多地变成了一种精确营销。网络经济利用用户的个人信息创造了巨大的财富。

2009 年《华尔街日报》引用的一项广告行业研究表明，无目的性的在线广告创造的价值是每千次点击量 1.98 美元，而有目的性的在线广告每千次点击量创造的价值是 4.12 美元（Jeffrey, 2011）。过去我们衡量网站成功与否的标准是计算浏览量，而如今我们更多地把它们看成是社交网络，即这些网站对用户的了解程度。这也解释了为什么沃尔玛公司会斥资 3 亿美元，收购提供社交媒体内容过滤及分类平台的 Kosmix 公司。Kosmix 团队的主要优势在于对社交网络内容和信息的分类、过滤以及优化搜索。在 Kosmix 上，如果用户搜索 Shoes，系统则会条理分明地显示来自社交购物网站 Kaboodle 的热门列表、Stylehive 的达人们的收藏标签、Youtube 的视频、Google 的图片以及来自 Twitter 的讨论，这显然区别于传统搜索引擎主题式的陈列。

Kosmix 令沃尔玛动心的一个大背景是，如今，社交网络对于零售业的重塑正在成为现实，而电子商务则是可以感受到这种变化的桥头堡。最明显的例子便是由社交网络跳转到电子商务网站的访问量增长迅速，对于电子商务而言，

访问量便意味着利益。

其他的零售商以及 Facebook、雅虎等网络巨擘，正在使用另一种开源云计算技术 Cloudera 来整理过去数年间存储的数量庞大的行为信息，借以寻找只有电脑才能分析计算出的行为模式。以这些方式产生的智能可以帮助 Zynga[①] 等社交游戏公司设计出更好的游戏，或为各个不同行业的品牌提供更好的广告创意。如果在线广告具备适当的目的性，它就有可能成为重要的信息。

潜在的利益帮助解释了各种数据交换、数据集市、预测分析引擎和其他中介产品的不断增加。这同时也解释了为什么谷歌、脸谱（Facebook）和 Zynga 以及其他许多公司，正在想方设法收集更多的用户信息。脸谱提供了一个例子，证明了这种追踪用户信息做法的广泛性。脸谱上的“喜欢”按钮看上去无伤大雅，点击“喜欢”，你就可以立刻分享你和朋友们都喜欢的信息；然而，如果你在登录脸谱账号的情况下访问了带有“喜欢”键的网页，脸谱就可以追踪你在这个网页上所看的内容及所做的事情。这难免会使人有些不安。

5.1.2　大数据对公众隐私与信息安全的威胁

互联网给人们带来海量的信息，提供了大量的机会，也进一步成为推动企业发展的重要手段。但是，在互联网飞速发展进步的同时，个人信息安全也受到了严重的威胁。

2011 年 4 月初，全球最大的电子邮件营销公司艾司隆（Epsilon）发生了史上最严重的黑客入侵事件，导致许多主要的企业客户名单以及电子邮件地址因此外泄，受害企业包括了摩根大通、第一资本集团、万豪饭店、美国银行、花旗银行、沃尔格林药妆连锁店及电视购物网络等。而就在不到一个月时间的同年 4 月底，索尼公司遭到黑客攻击，泄露了一亿份账户资料，将其 Play Station 网络[②] 和 Qriocity[③] 流媒体服务关闭了将近一个月。索尼公司因此花费了约 1.71 亿美元来弥补这个损失。

然而，黑客并不是互联网时代人们隐私和信息安全的唯一威胁者。同样在

① Zynga 是一个社交游戏公司，于 2007 年 6 月成立。Zynga 开发的游戏多半是网页游戏，并发布于 Facebook 以及 MySpace 一类的社交网站。公司的总部在美国旧金山。

② PlayStation 是日本索尼公司旗下的新力电脑娱乐 SCEI 家用电视游戏机，现已成为最出名的家游产品之一。玩家只要把 PlayStation 主机连接上网，便可以即时下载免费游戏、参加网上对战，体验 PlayStation 网络的强大功能。

③ 索尼公司推出的线上云端影音串流服务，使用户能订阅音乐推送到一切索尼设备上，包括各种播放器和索尼的 Bravia 电视等可上网连线的装置都可以使用这个平台。

那个 4 月,《华尔街日报》报道（Julia et al.，2011）说，其安全分析师发现苹果 iPhone 和使用谷歌安卓操作系统的智能手机会自动收集用户行踪信息，而且苹果手机在定位功能被关闭后仍会继续收集和保存用户位置信息。虽然苹果公司发表了否认声明,并宣布将发布软件升级程序来修补技术漏洞,但这一“跟踪定位”事件在美国和其他国家受到了广泛关注，引发各界对移动设备与个人隐私保护问题的新一轮的讨论和思考。

《华尔街日报》2010 年 7 月的一篇文章《个人隐私：互联网新金矿》中报道了一名美国女性令人不安的经历（Julia, 2010）：

在艾希莉·海耶斯·比蒂 (Ashley Hayes-Beaty) 的电脑里，一个小小的文件正在帮助收集关于她的各种个人信息。最终这些信息将以 1/10 美分的价格被出售。这一文件包含着一个简单的代码——4c812db292272995e5416a323e79bd37。懂行的人会知道，这个代码说明艾希莉是一名来自田纳西州纳什维尔的 26 岁女性，最喜欢的电影是《公主新娘》、《初恋五十次》、《对面恶女看过来》，也知道她最喜欢的电视剧是《欲望都市》，更知道她喜欢浏览娱乐新闻，喜欢各种问答。

艾希莉正在一家纽约公司 Lotame 的监控之下。这家公司使用一种基于网络标签和信标的复杂软件，来捕捉人们在网站上输入的文字，如对电影的评论，也可以追踪他们感兴趣的网页内容究竟是哪些。Lotame 最终会将很多这样的个人信息打包出售，卖给那些正在寻找潜在消费者的企业。比如，艾希莉的资料可以纳入电影爱好者的包裹，价格是每 1000 人 1 美元。当然，也可以更加详细地定制，将这些资料按照各种不同的方式进行逐层细化，比如艾希莉可以被界定为“《初恋五十次》的 26 岁南方影迷”。

《华尔街日报》的调查发现，目前互联网上成长最快速的生意之一就是监视互联网用户。全美最大的 50 家网站在访问者的电脑上平均安装了 64 种追踪技术,通常都没有任何警告。大约 10 多家网站所安装的技术甚至超过了上百种。此外，追踪技术正在变得越来越精巧，植入程度也越来越深。过去，监控一般都被局限在记录用户访问网站情况的 Cookies 之中，而《华尔街日报》却发现，新的工具完全可以在人们浏览网页和操作的时候进行实时扫描，然后立即对所在地、收入、购物兴趣，甚至医疗条件等因素进行评估。一部分工具甚至可以在用户试图删除它们的时候悄悄地进行自我复制。

在这个大数据时代，我们的线上生活几乎都是可以被追踪的，甚至线下生活也可以被追踪（Jeffrey, 2011）。就拿一家座落在美国硅谷的新兴弹性社交网

络公司 Color 来说吧。这家公司旨在利用手机设备里的 GPS 定位功能，结合内置的陀螺仪和加速计来解析手机用户所拍的照片流并据此定位用户所处的位置。通过观察用户通过 Color 的软件所分享的照片，分析图片涉及的内容，加上手机麦克风所搜集的环境声音，Color 不仅可以显示用户所在的位置，而且能反映用户正和谁在一起。这种服务不仅对于那些对手机社交网络感兴趣的用户而言十分具有吸引力，而且也吸引了许多狂热的技术爱好者。

Color 公司的做法说明了一个越来越突出的事实：企业正在日渐掌握新的方法来捕获关于消费者的信息。如今，它们已经拥有了使数量巨大的非结构化松散数据变得有意义的技术，如自然语言处理、机器学习，以及诸如分布式计算（Hadoop）这样的软件架构，可以处理大量的同步网络搜索请求信息的分析。网络搜索信息这种杂乱无章的数据，早已被归入数据仓库，如今已成为数据挖掘的主要对象。社交网络所生成的信息也是如此，主要包括个人资料、发帖和日志等。这些信息的数量令人咋舌，国际数据公司（IDC）的一份报告估测，2009 年存储的信息总量达 0.8ZB，相当于 8000 亿 GB；国际数据公司预测，到 2020 年，全球存储的数据信息将达到 35ZB，这其中的大部分都是客户信息（Jeffrey, 2011）。随着数据存储量的增加，从中通过分析而得出推论和预测的做法将越来越普遍、越来越熟练。

然而，这些数据的使用是否应该得到用户的授权？这些信息是否会遭到滥用？用户是否会因为信息泄露而遭到骚扰？个人信息是否会遭到断章取义的理解？这些都涉及公众的隐私与信息安全，必须得到重视。

5.1.3　保护公众隐私与信息安全的对策

怪不得如今有人呼吁企业要设立隐私主管、安全主管、数据主管等职位，美国和欧洲的立法者也正在考虑采取各种方式保护公众的隐私（Jeffrey, 2011）。在 2011 年美国共和党和民主党的一份联合提案中，参议院约翰 • 麦凯恩和约翰 • 克里共同提出了《消费者隐私权利法案》（*Consumer Privacy Bill of Rights*）的议案，其目的是在部分程度上限制互联网公司对消费者数据的使用。参议院杰 • 洛克菲勒也在 2011 年提出了独立议案，即《网络不跟踪法案》（*Do-Not-Track Online*）的议案。

在过去几个月里，美国政府已经采取了一系列重大措施，使广大消费者能够对自己的个人在线信息拥有更强的控制权。2012 年 2 月下旬，奥巴马政府

公布了《消费者隐私权利法案》[①]。数周之后，美国联邦贸易委员会（FTC）发布了有关消费者隐私权利保护的最终报告，该报告敦促各家私营公司采取自我监管的做法。值得称道的是，美国的私营部门正在日益努力增强自我监管措施，其中包括那些来自诸如网络广告倡议组织、互动广告局、数字广告联盟等组织发起的措施。许多广告技术公司已经建立了自己的“不跟踪”机制，消费者可以通过点击在广告旁一起出现的“广告选择”（AdChoices）图标，选择不参与（opt out）接受针对性广告的服务活动。

在私营部门领域，美国行业自律组织数字广告联盟（Digital Advertising Alliance）走在了官方政策出台的前面。数字广告联盟于 2009 年由几大媒体协会共同组成，包括美国广告代理协会、美国广告联盟、美国广告主协会、美国直销协会、美国互动广告局、网络广告促进会等。2010 年底，数字广告联盟推出了它自己设计的隐私保护框架“在线行为广告自我监管项目”，以保证消费者信息的安全。这个项目倡议广告平台、广告客户、广告公司以及网民共同营造一个可信的、规范的、有统一反馈机制的广告环境。承诺参与该项目的有谷歌、美国在线、雅虎等美国主流广告发布平台，微软、戴尔、通用汽车公司等大型广告客户也对这个项目给予了大力支持。这种做法是媒体合作与行业自律的产物，而消费者就是最大的受益人群。

欧盟数据保护工作组也正在致力于解决同样的问题。欧盟数据保护工作组曾在 2009 年分别致信谷歌、微软和雅虎三大搜索引擎巨头，认为搜索引擎服务商保存用户搜索记录时间超过 6 个月的理由并不成立，因此要求这三大搜索引擎商必须缩短用户搜索信息的保留时间。

大数据潜在的黑暗面意味着我们需要道德准则来规范大数据的使用。杰弗里·雷波特在《大数据需要道德准则》一文中为企业提出了四条建立使用大数据道德准则的建议（Jeffrey, 2011）。

1. 做法公开

收集数据时，要让用户及时知道。这样的公开做法可以应对隐藏文件和非授权追踪而造成的问题。让用户知道公司对于他们信息的掌握程度，有助于建立用户与公司之间的信任，谷歌公司已经开始这么做了。如果你想知道谷歌了解了你的哪些信息，可以点击 www.Google.com/ads/preferences 了解你的广告偏好设置，页面会显示你的兴趣已与你浏览器中存储的某个广告 Cookie 相关

① 该法案为如何保护用户隐私设定了 7 项原则，其中包括网络用户有权控制哪些个人数据可以被收集和使用，有权得到易于理解的有关隐私和安全方面的信息，个人信息被收集、使用、披露的方式必须与用户提供这些信息的背景相一致，企业必须负责任地使用用户信息等。

联，如果不希望谷歌存储你的兴趣，则可以在下方选择“停用”。

2. 设置简单

为了防止公众被蒙在鼓里从而造成信息的滥用，企业应该给用户机会，让他们自己去弄明白到底想要什么程度的隐私。因此，许多网络企业都设有隐私政策。脸谱的隐私政策共有 5830 个英文单词，比美国《宪法》的 4543 个词（不包括修正案）还多。但这只是冰山一角。企业应该试着修改隐私政策，使其简明易懂、一目了然。

3. 从设计着手保护隐私

有些人认为做法公开和设置简单都不足以保证用户的网络隐私。加拿大安大略省的信息与隐私专员安·卡沃基安提出了“设计隐私”的概念，倡导各机构组织将隐私保护加入所有活动和产品之中。她认为，没有人愿意仔细看完隐私政策，从一开始就为保护隐私投入少量资金的话，能防止数据泄露和品牌形象受损，还能省下不少冤枉钱。但这并不意味着网络和移动产品不会收集用户信息，只是说这些企业和组织会从一开始就将保护用户隐私作为一项基本准则。微软公司 2006 年发布的一份名为《研发软件产品和服务的隐私准则》报告就明确地体现了这一点。微软的最新浏览器 IE9 使用户可以通过开启设置来阻挡第三方广告。目前在隐私设计方面较为成功的典型是谷歌的新社交网络 Google+。此前谷歌推出的产品 Google Buzz 自动根据 Gmail 用户的联系人名单创建社交网络，但联系人信息理应是私密的，这种做法违背了最基本的隐私原则。后来，谷歌在新社交产品中将隐私保护作为基石，所有联系人都置于非公开的“圈子”（如“朋友”、“同事”和“家人”）中。每当用户进行分享时，他们需要选定分享给哪个圈子。

4. 交换价值

当你走进一家星巴克咖啡店，如果服务员记得你的名字和你喜欢喝的咖啡口味，你很有可能会觉得受宠若惊。同样的事情也会发生在网上：一家服务提供商对你的了解越多，你就越有可能喜欢它的服务。彻底的公开透明可以使数字商业更便捷地向用户展示它们可以为用户提供的服务，以作为对用户分享个人信息的交换。收费视频网站网飞公司（Netflix）实践了这一做法。该公司举办了一次公共竞赛，向第三方研发商提供 100 万美元的奖金鼓励他们研发最有效的电影推荐引擎。网飞公司使用用户的电影浏览记录来提供针对性不断提高、更加有用的观影推荐，这一点是众所周知的。

这些原则并不详尽，只是列出了企业对大数据价值的看法及缓和其风险的做法。采用这些原则也可以帮助企业走在决策制定者试图管理数字经济做法的

前面。其实，关于使用大数据的最重要的黄金法则还是那句老话：己所不欲，勿施于人。这种想法和做法可以帮助我们创造期望的，也是应有的数字世界。

附：谷歌隐私权原则

谷歌公司的创意和产品经常超越现有技术，从而推动技术的不断进步。作为一家负责任的公司，我们努力确保在进行任何创新的同时，都能为用户提供相应级别的隐私权和安全性。全公司上下在制定决策时都会以隐私权原则为指导。这样，我们就能在完成“整合全球信息，使人人皆可访问并从中受益”这一长期使命的过程中，让用户得到保护并掌握更多信息。

利用收集的信息为用户提供有价值的产品和服务

谷歌十大信条的第一条便是“以用户为中心，其他一切水到渠成”。用户与我们分享信息后，我们可以反过来利用这些信息为用户提供有价值的服务和产品。我们相信，以用户为中心可以促使我们开发出各种产品以及有助于强化隐私权的功能，而正是这些为我们带来了创新和忠实的在线用户群。

我们会研究人们在搜索中常犯的输入错误和拼写错误，以帮助您更快、更准确地获取搜索结果。因此，如果您输入“周结论”，我们就会推测您可能要搜索的是“周杰伦”。

开发符合隐私权标准和隐私权惯例的产品

我们的目标是引领技术潮流，这包括开发各种工具，帮助用户简单方便地管理自己的个人信息，同时不会对我们所重视的用户体验造成任何不利影响。我们遵守各种隐私权法律；此外，我们还通过内部的工作以及与监管机构和行业合作伙伴之间的合作，制定并实施严格的隐私权标准。

我们设计了带有圈子功能的Google+，以便让您轻松地与不同的对象分享各种内容。您可以使用这个产品将朋友们放到一个圈子中，将家人放到另一个圈子中，而将老板单独放到一个圈子中，就跟现实生活中一样。

将个人信息的收集透明化

我们会尽一切努力向用户显示我们用于自定义服务的信息。我们希望尽可能地让用户知晓我们所收集的个人用户信息，并说明我们如何使用这些信息提供服务。

您可以访问谷歌信息中心了解谷歌知道您的哪些信息，那里会显示您的谷歌账户中存储的信息（例如您博客中的最新博文或者是您上传的照片），还可让您从一处集中更改多个谷歌产品的隐私设置。

为用户提供有意义的选择，保护他们的隐私权

每个人对于隐私权的关注点和需求各有不同。为了能让所有用户均享受到

最好的服务，谷歌力求为用户提供多种有意义且体贴细致的选择，帮助他们决定如何让谷歌使用其个人信息。我们认为，个人信息是不应受到制约的，所以我们致力于开发可让用户将自己的个人信息导出到其他服务中的产品。我们不会出售用户的个人信息。

借助我们的隐私权工具，您可以对自己的计算机与谷歌之间的搜索流量进行加密、隐身浏览互联网、删除搜索记录、使用我们的数据备份功能轻松将数据移出谷歌产品，您还可以实现更多其他功能。

安全保护我们掌握的信息

我们将责无旁贷地保护用户提供给我们的数据。我们非常重视安全性问题，并与广大用户、开发人员和外部安全专家通力合作，共同营造更加安全可靠的互联网环境。

我们从一开始就在产品中注入了安全性和可恢复性的设计理念。我们的自动扫描器每天都在使用各种数据保护数以百万计的用户免受恶意软件、网上诱骗、欺诈和垃圾邮件的侵害。

5.2 信息选择与决策制定

5.2.1 大数据不等于全数据

现今对于大数据的研究，很多都是从计算的角度出发的，但同时我们必须了解到，这些数据都是关于人的，因此要考虑到纯计算带来的局限性（Danah, 2010）。此外，大数据数量庞大，但数量不等于质量，质量比数量更重要。大数据存在一定的局限性，其中之一就是样本选择问题。样本选择对所有的社会科学学科都至关重要，选中的样本和数据影响着研究结论的得出。

从理想的方法论上来讲，研究者希望能取得所有人口的相关数据，以更好地抉择应该如何取样。如果有这样的一个完整数据库，计算频数的学者就可以轻易地取得有代表性的随机样本。同样，想要了解多样性的学者能够取得异常值。从历史上来看，能够取得有关任意一方面的全部数据以使研究者得出结论的最好近似值，这是非常罕有的。

大数据的出现让我们看到了完整数据库存在的可能性。但是，大数据并不总是完整的。推特拥有整个推特上信息的数据库，但大多数研究者无法获得推特上的所有数据，至多只能看到所有公开的推特信息。通常这些信息是公开话题下能被搜索到的，而这些有时也不是随机显示的。

数量大与数量全并不是一回事。如果要研究推特上关于某个话题的频数，而包含问题词汇的推特信息无法被搜索到，那么基于这一数据样本而做出的分

析就是有问题的。

取样需要研究者在做研究的每一阶段都要摒弃自己的偏见。某些类型的人有没有受到同等的重视？如果没有的话，意味着什么？假设研究者能够取得推特上所有公开推特的信息，如果对这些信息进行随机取样，那么就意味着研究者并没有对用户账号进行随机取样。这是因为，并不是所有用户账号发布推特信息的频数都是一样的，有些人发得勤、发得多，有些人则发得少。这样一来，随机取样中那些发布信息勤快的账号就会提供更多的推特信息。再假设研究者可以从所有推特账号中随机取样。但这样一来，研究者只是在对推特账号进行随机抽样，并没有对推特用户进行抽样，因为有些用户拥有多个推特账号，有些用户有账号但不发推特信息，只是围观，而有些人没有账号却经常阅读推特上的信息。

这就好比当个人拥有多张手机卡时，汇合每张手机卡上的数据并把它们归到同一个人身上几乎是不可能的。数据与人口统计指标联系起来时，才能发挥最大用途，因为此时的数据可以反映某一部分人群的习惯。因此我们需要改善将订阅服务与人口信息相联系的方式，从而确保移动设备产生的数据能最大程度地体现个人化。

因此研究者必须对数据集有全面的了解。数以百万计的数据并不意味着它们就是随机的、有代表性的。要用数据来进行解释分析，就必须清楚地知道数据的出处。

5.2.2　大数据不等于真数据

因为大数据的“大”，很多与大数据打交道的人就认为它是最好的数据。大数据的价值显而易见，但也存在着局限性：它只能揭示和解释某些事情。如果研究者认为大数据能说明的事情比它实际能做到的多，那就是一件危险的事情。

就拿社会网络的研究来说吧。不同学科的学者都在用各种方法和分析手段来研究社会网络。但我们并不能说从脸谱上或手机记录中获得的数据就比社会学家以其他手段获得的信息更加准确。它们都是极其有用的网络，但存在着不同之处。

从历史上来讲，只有社会学家对研究社会网络感兴趣，通过调查、访问、观察和实验等方式收集有关社会网络的数据。通过这些数据，社会学家将有关个人社会网络的信息进行理论化提炼。争议的焦点在于应该如何测量分析个人网络，某些研究方法是否会导致偏见以及该如何解释这些偏见。

大数据引入了两种新近流行的由数据追踪而产生的社会网络类型，即表

达型社会网络与行为型社会网络。表达型社会网络指的是社会网络上公开显示的结果，如脸谱上用户公开的好友名单；行为型社会网络指的是沟通交流模式。这两种社会网络与社会学家所讲的传统意义上测量和理论化的网络都不一样。

脸谱上许多人所列的好友是不相识或者不喜欢的人。因此，分析脸谱上的好友名单不能等同于分析了一个人的社会网络。某个社交网站的用户在网站上所列的特别好友可能不是现实生活最亲近的朋友，做出这样的选择可能是出于一些其他因素的考虑。因此，通过网络联系的频数和公开显示的信息来判别人际关系的亲疏，是存在局限性的。

最爱大数据的是市场营销人员，而最容易误读大数据的也是他们。这是因为市场营销人员一般认为"是什么"能够回答"为什么"的问题。大数据显示的只是已经呈现的表面现象和结果，如果不通过询问和访谈，可能就无法得知数据背后人们真正的想法和意图。

分析人们的行为模式和互动模式是一项极其重要的研究工作，但这只是理解社会动态的第一步。一味地分析数据只能帮助研究者看到现象和结果，如果不与人们交谈，就无法了解人们行为的背后原因。"是什么"和"为什么"是两个不同的问题。如果在知道了"是什么"的基础上，通过自己的猜想来说明"为什么"，这在方法上就是错误的。因此，通过大数据来进行解释和分析，并不是一件容易的事。

不仅仅是研究者，大多数人都没有完全理解在回答"为什么"这个问题时，"是什么"其实可以有不同的诠释。Cobot 软件公司曾通过一款 LambdaMOO 的社区网络产品收集其网民使用者的数据。由于该公司没有利用这些数据，网民觉得不安，就要求公司使这些数据物尽其用。因此，Cobot 公司重新编写了一个程序，用户可以询问有关公司所收集数据的问题。不久后，用户就开始询问他们使用这款产品时和谁的对话最多，接着开始问他们的朋友最经常对话的对象是谁。可想而知，当甲得知，他最经常对话的乙却与丙联络最频繁时，他怒火中烧，再也不和乙说话了。这一网络社区中的许多人将这些产品使用的行为信息当作关系亲疏的指示器。这个网络社区最后分崩瓦解。

鉴于大数据所存在的局限性，对大数据的误读频频发生。通过大量数据和精细的测量，统计学家和计算机科学家注意到，"错误发现"的威胁正在上升。斯坦福大学的统计学教授特雷夫 • 哈斯蒂（Trevor Hastie）说，在数据的干草堆中捞到有意义的"针"，其困难就是"许多干草看起来也像针"（Lohr, 2012）。

大数据也为统计诡计和有偏见的事实发现型研究提供了许多原材料。它

为一个旧把戏套上了高科技的面具，即我知道事实，如今就只需要发现它们。数据通过计算机和数学模型来进行梳理，从而被理解。这些模型就像文学中的比喻一样，简单化地解释信息。它们有助于人们理解信息，但也有其局限。模型可能根据网络搜索检索出数据的相互关系，得出不公正或歧视性的统计性推断。

因此，企业、组织和政府在利用大数据时要充分考虑其局限性，在收集信息和进行分析时注重研究方法的运用，使大数据能够尽可能地反映事实。

5.3　大数据与非传统安全

5.3.1　网络恐怖主义与信息战的威胁

在机械化战争时代，国家众多目标中直接面临挑战的是军队。但在信息时代，国家安全环境发生了质的变化。无论在战时还是在平时，一国的各种信息设施和重要机构等都可能成为打击目标，而且保护它们免受攻击已超出了军事职权和能力的范围。决策的不可靠性、信息自身的不安全性、网络的脆弱性、攻击者数量的激增、军事战略作用的下降和地理作用的消失等都使国家安全受到了严峻的挑战。此外，网络化的国家在石油和天然气管道、水、电力、交通、银行、金融、商业和军事等方面都依赖信息网络控制系统，因而容易遭受信息武器的攻击。在信息时代，信息攻击可从任何地方发起，可在瞬间穿过任何自然障碍，从而使地理作用降到有史以来的最低点，也使任何国家无法再享受到天然的“安全保障剩余”[①]（陈效卫，2001）。

22 年前，“信息战之父”沈伟光[②]就出版了世界上首部《信息战》学术专著，提出了信息战的概念。而在如今大数据时代到来之际，丰富又多样的数据来源无疑又为信息战提供了“火力”支持。大数据将对各国具有重要的战略安全意义。

此外，大数据也将为网络恐怖主义提供新的资源支持。庞大海量的大数据涉及的方面之广，将有可能使网络恐怖主义的势力侵入人们生活的方方面面。

---

① “安全保障剩余”：国际政治周期性规律的创立者、美国著名学者莫德尔斯基的观点，认为一个国家要成为国际体系中的真正强国，一个不可或缺的条件是享 有“安全保障剩余”（Security Surplus），即居有岛国或半岛国的地缘位置，使自身享有一种进可攻、退可守的战略自主性。

② 沈伟光，1959 年 7 月 23 日出生，浙江杭州人。未来学家，信息战专家。美国人称他为“信息战之父”。现在在浙江省档案局工作。1985 年开始研究并提出信息战概念，1987 年 4 月 17 日《解放军报》以《信息战的崛起》为题报道了他的研究情况；1990 年 3 月在浙江大学出版社出版了世界上第一部《信息战》专著，又提出信息边疆、信息化战争、信息化军队等新战争概念。

一般而言，网络恐怖主义指恐怖主义与网络空间的结合，是一种由国家或非国家主使的，针对信息、计算机程序和数据以及网络系统带有明确政治目的的攻击行动。具体来说，网络恐怖活动的行为主体是电脑网络黑客，攻击目标是一国或数国的计算机与信息网络系统，其手段和方式是使用针对计算机操作系统的漏洞和网络软件的缺陷开发出来的黑客程序软件。它通过威胁、攻击以及破坏和瘫痪某国的民用或军事基础设施，制造心理恐慌，造成财富损失，从而达到某种政治与社会目的（俞晓秋，2002）。

“9·11”事件后，一些伊斯兰黑客组织像穆斯林游击战士为报复美国黑客的攻击，攻击了美国海洋及大气局网站，并在其网页上留下恐吓字句，威胁称如果美国不停止打击阿富汗以及基地组织，他们会把手上的美政府机密资料交给基地组织。他们还攻击美国国家卫生研究所全国人类基因组织机构的服务器，涂改了网页，贴上了沙特阿拉伯国旗并留下两条乌都尔文标语“真主伟大至极”和“美国人准备受死吧”。因此，社会价值感的扭曲以及无政府主义思想的膨胀，将导致黑客实施国家规模或国际规模的恐怖袭击，从而蜕变成网络恐怖主义者（刘强，2004）。

网络恐怖主义比传统意义上的恐怖主义活动更加防不胜防。如何保证数据的安全，将是大数据时代的一项严峻挑战。

5.3.2　案例：美国的对策

为了更好地利用信息技术来反对恐怖主义的袭击，美国联邦政府正在研究并实施一些新方法，利用海量的、以商业手段收集的个人信息数据库来为提高国家安全服务。这些信息库几乎包含了各个行业，包括保险信息、旅游信息、金融数据、零售记录，以及法庭文件、证书和房产证明等政府部门资料。这一趋势早在 2001 年“9·11”事件发生前就已经产生，但从那之后不断增强，新的数据环境已经产生了两大前所未有的特征，即来源于私人部门的、可用的个人化识别信息具有深度和广度，同时用于分析这些数据的分布形势与意义的能力也在不断提高（James et al.，2004）。

2012 年 3 月 29 日，美国联邦政府宣布了《大数据研究和发展倡议》（*Big Data Research and Development Initiative*），斥资 2 亿美元投入大数据研究领域，以加强政府各个部门、研究机构和其他组织从大量复杂的数据中提取、分析重要信息的能力。这一倡议涉及美国联邦政府的六个部门，分别是美国国家科学基金、美国国家卫生研究院、美国能源部、美国国防部、美国国防部高级研究计划局和美国地质勘探局。这些部门将大力推动和改善与大数据相关的收集、

组织和分析工具及技术的研发和使用，力图在科学发现、环境保护和生物医药研究、教育、国家安全及战争策略等领域利用大数据能力取得突破。

奥巴马政府宣布用 2 亿美元投资大数据领域，体现了大数据对于国家发展和国家安全的重要性。大数据作为一种新型的经济资产，同时具备安全和战略意义。大数据技术领域的竞争，事关国家安全和未来。

目前，美国在研究与利用大数据方面走在世界前列，英国紧随其后，而大数据在世界的其他国家还是一个新兴概念，因此相对而言研究和利用得还比较少（Mac，2012）。然而，随着大数据重要性的逐渐体现，不仅商业领域将更多地利用大数据，各国政府也会更加重视大数据，进而将这种新型资产用于提高国家安全这一重要领域。

附：美国政府大数据计划①

美国联邦政府为了应对大数据革命所带来的机遇，制订相关计划以推进有关研究机构在大数据方面进一步实现科学发现，并开展创新研究。

国防部（DOD）

国防部高级研究计划局（DARPA）

（1）多尺度异常检测项目（ADAMS）解决大规模数据集的异常检测和特征化问题。项目中对异常数据的检测是指对现实世界环境中各种可操作的信息数据及线索的收集。最初的 ADAMS 应用程序进行的是内部威胁的检测，即在日常网络活动环境中检测单独的异常行动。

（2）网络内部威胁计划（CINDER），旨在开发新的方法来检测军事计算机网络与网络间谍活动。作为一种揭露隐藏操作的手段，CINDER 将适用于对不同类型对手的活动统一成“规范”的内部网络活动，目的是提高对网络威胁检测的准确性和速度。

（3）洞察识别项目（Insight）计划主要解决目前情报、监视和侦察系统的不足，进行自动化和人机集成推理，使得能够提前对时间敏感的更大潜在威胁进行分析。该计划旨在开发出资源管理系统，通过分析图像和非图像的传感器信息和其他来源的信息，进行网络威胁的自动识别和非常规的战争行为。

（4）机器阅读项目（Machine Reading）旨在实现人工智能的应用及在发展学习系统的过程中对自然文本进行知识插入，而不是依靠昂贵和费时的知识表示目前的进程，并需要专家和相关知识工程师所给出的语义来表达信息

① 资源翻译自：Big Data Fact Sheet, Executive Office of the President, March 29, 2012, http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf。

的含义。

（5）思想之眼项目（Mind's Eye）旨在为机器建立视觉的智能。传统的机器视觉研究的对象选取广泛的物体来描述一个场景的属性名词，而思想之眼旨在增加在这些场景的动作认识和推理需要的知觉认知基础。总之，这些技术可以建立一个更完整的视觉智能效果。

（6）以任务为导向的反应云项目（Mission-oriented Resilient Clouds）通过技术进行检测、诊断并对攻击作出响应，有效地建立了“社区卫生服务系统”的云，以解决云计算固有的安全挑战。该方案还旨在开发新技术，使云应用和基础设施受到攻击时能够继续运行。只要整体能够有效地运行和保存，可以允许个别主机和任务损失。

（7）对加密数据的编程计算（PROCEED）的研究工作旨在开发实用的方法与相关现代化的计算编程语言，使数据加密时仍然能使用云计算环境，以克服信息安全的重大挑战。使用户能够在不需首次解密的情况下操纵加密的数据，它将使得对手拦截信息更加困难。

（8）在视频和图像的检索与分析工具（VIRAT）计划旨在开发一个系统能够利用军事图像分析员收集的数据进行大规模的军事图像分析。VIRAT 如果成功，将使分析师在相关活动发生时建立警报。VIRAT 还计划开发工具，能够以更加高的精准率和召回率从大量视频库里实现视频内容的检索。

（9）XDATA 项目计划旨在开发用于分析大量的半结构化和非结构化数据的计算技术及软件工具。最核心的挑战是，可伸缩的算法在分布式数据存储应用、如何使人机交互工具能够有效地迅速定制不同的任务，以方便对不同数据进行视觉化处理。对开源软件工具包的灵活使用，使得能够处理大量国防应用中的数据。

国土安全部（DHS）

可视化及数据分析卓越研究中心（CVADA）是由罗格斯大学与普渡大学（以及另外三所大学）的研究人员共同合作建立的，主要从事对大量异构数据进行研究，使相关人员可以发现人为或自然灾害、恐怖事件、需要执法的边境安全问题、网络威胁爆炸物等。

能源部（DOE）

科学办公室

（1）高级科学计算研究办公室（ASCR）主要负责数据管理、可视化和数据分析，包括数字化保存和社区访问等。套件程序里包括广泛使用数据管理的技术，如开普勒科学的工作流程系统、存储资源管理标准；各种数据存储

管理技术，如 BeSTman、大容量数据移动器和适应式的 IO 系统（ADIOS）；FastBit 数据索引技术（雅虎使用）和两个主要的科学可视化工具——ParaView 和 VisIt。

（2）高性能存储系统（HPSS）是对磁盘和磁带系统上 PB 级数据进行管理的数据管理软件。由美国能源部和 IBM 开发的 HPSS 在世界各地的大学和实验室的使用，用在数字图书馆、国防应用和包括纳米技术、基因组学、化学、磁共振成像、核物理、计算流体力学、气候在内的一系列学科，以及诺斯罗普•格鲁门公司，美国国家航空航天局（NASA）和美国国会图书馆。

（3）千万亿次数据数学分析主要是对千万亿次的数据分析处理从庞大的科学数据集提取信息，发现其主要特征，并理解其间的关系。研究领域包括机器学习，数据流的实时分析，非线性随机的数据缩减技术和可扩展的统计分析技术，广泛适应于从能源部到电网，包括宇宙学和天气数据、传感器数据等。

（4）下一代网络方案支持工具使得合作研究在重大发现方面能够有所作为，包括 2001 年的 Globus 中间件项目大量数据的移动和使用、2003 年的 GridFTP 的数据传输协议、2007 年的地球系统网格（ESG）的工具。今天的 GridFTP 的服务器开放科学网格、地球系统网格和生物社区的科学数据每月超过 1 PB 的移动。Globus 中间件也被得克萨斯大学、软件公司、石油公司利用并一起合作，培养学生能够使用先进的石油工程方法和集成的工作流程。

基础能源科学办公室（BES）

这一办公室的科学用户设施支持旨在协助用户数据管理和分析大数据，可每天从一个单一的实验数据大容量化（1012 字节）努力。例如，加速数据采集，处理和分析（ADARA）项目解决了数据的散裂中子源（SNS）的数据系统提供实时分析，实验控制的工作流程需要，以及已经建立 X 射线影像资料库，以最大限度地提高数据的可用性和更有效地利用同步加速器光源。

在 2011 年 10 月，由生物工程学会和 ASCR 的基础能源科学的数据和通信研讨会将确定实验数据的需求，这可能会影响到科学发现。

（1）生物和环境研究计划（BER）大气辐射测量（ARM）气候研究设施是一个多平台的科学用户设施，提供重要的大气现象的精确观测研究，大气过程认识的进步需要国际社会的基础设施和气候模型。ARM 的数据是可以进行应用的，并以其作为文章发表在一个超过 100 年历史的杂志。正在处理收集和展示的高时间分辨率和光谱信息，可应对与数百份文书相关的挑战，以满足用户的需求。

（2）系统生物学知识库（Kbase）是一个社区驱动的软件框架，对微生物、植物和环境条件下的生物群落功能的数据驱动的预测。这是一个开放式的设计

与开发，以提高算法的开发和部署效率，并增加从异构数据源的实验数据的获取和集成。Kbase 不是一个典型的数据库，而是一种手段，以解释缺少的信息成为实验设计预测工具。

聚变能源科学办公室（FES）

通过 FES 和高级科学计算研究（ASCR）办公室高级计算合作的（SciDAC）科学发现在聚变能的科学计算和实验研究大数据存在的挑战。ASCR-FES 开发的数据管理技术，包括高性能的输入/输出系统，先进的科学工作流程和出处框架，可视化技术解决独特的融合需求，已经吸引了欧洲一体化建模的努力和国际热核实验堆，一个国际核聚变研究和工程项目的关注。

高能物理办公室（HEP）

高能物理计算计划经过了全球数百名科学家的努力，支持大量的分析研究，复杂的实验数据集，以及大量的模拟数据。协作企业进行大数据管理，包括生产和分布式分析 PanDA（产品分布式分析）工作量管理系统及 XRootD，一种高性能、快速、可扩展访问多种数据存储库的容错软件。

核物理办公室（NP）

美国核数据计划（USNDP）是一个多方面努力，涉及 7 个国家实验室和 2 所大学的项目，提供跨越多个领域，包括核物理、编译和交叉检查，对所有原子核的重要性质的相关实验结果、维护和广泛使用的专用数据库。

科学和技术信息办公室（OSTI）

OSTI 是唯一的 DataCite 美国联邦机构成员（全球领先的财团科学和技术信息的组织）中发挥了关键作用，在塑造实践的政策和技术实现数据的引用，这使得可以跟踪数据的影响，使有效的重用和数据核查与学术结构的表彰及奖励数据生产商可设立。

退伍军人管理部（VA）

（1）医疗保健信息研究所（CHIR）开发自然语言处理（NPL）工具，能够对在 VA 以文本形式存储的大量数据进行信息解锁。

（2）VA 正在努力通过保护作战人员使用文字处理算法捕获公共卫生事件（ProWatch），正在开发一个的生产透明、重复性好、可重复使用的各种安全相关的事件监控软件探测，以研究为基础的监控程序，能够跟踪、测量与军事部署相关的健康条件。

（3）AViVA 是 VA 的下一代就业人力资源系统，将业务应用程序和基于浏览器的用户界面分开的数据库。分析工具已经被建立在此基础上研究，最终决定在对病人进行支持。

（4）医学成果观察项目设计用来比较各种安全监测分析方法的有效性、可

行性和性能。

（5）企业数据仓库（CDW）是 VA 的项目，组织和管理从各种渠道传递的个人及群体的疾病和治疗的完整视图的数据。

（6）健康资料库是卫生保健提供者的数据格式规范的数据，尤其是 VA 和国防部之间，让 CDW 集成的数据。

（7）基因组信息系统综合科学（GenISIS）计划，通过个性化医疗，提高退伍军人的医疗保健。GenISIS 通过接触获得电子健康记录和遗传数据，可以跨 VA 进行临床试验、基因试验和成果研究。

（8）“百万美元老将计划”招募退伍军人自愿参加血液样本的基因分型和基因测序。这些基因样本支持 GenISIS，将用于了解个别老将的遗传疾病状态的健康记录。

（9）VA 的信息和计算基础设施为目前 VA 可用的大型数据集提供分析场所和工具，将促进 VA 任何网络之间的合作研究。

卫生和人类服务部（HHS）

疾病预防控制中心 (CDC)

（1）生物传感 2.0 是第一个考虑到区域和国家协调的情况下，通过互操作的网络系统对公众健康意识的可行性分析的系统，其建立在现有的国家和地方的能力之上。生物传感 2.0 移除许多单片物理结构相关的成本，同时还对最终用户透明的分布式系统方面，做出适当的分析和报告的数据访问。

（2）疾病预防控制中心的特别细菌学参考实验室（SBRL）的使用细菌和疫情 ID 网络生物学技术从有效、迅速爆发中检测未知的细菌病原体。谱系基因组学，比较整个基因组 DNA 序列的系统发育分析，将带来基于序列识别的概念，以全新的水平，在不久的将来对公众健康产生深远的影响。发展一个新的物种鉴定 SBRL 基因组管道，将允许多个分析一个新的或迅速崛起的病原体在几小时内进行，而不是数天或数周。

医疗保险和医疗补助服务中心 (CMS)

（1）正在开发基于 Hadoop 的数据库，以支持医疗保险和医疗补助项目的分析和报告。其主要目标之一是开发可支持、可持续、可扩展的系统设计，可在数据库一级进行数据积累，并补充现有的技术。

（2）正在评估使用 XML 数据库技术，以支持保险等事务密集型数据交换的环境，尤其是要支持资格筛查、报名等流程。XML 数据库可能可以容纳大数据表规模的数据，并对交互性能进行了优化。

（3）医疗保险和医疗补助服务中心与 Oak Ridge 国家实验室共同开展了一套试点项目，涉及数据可视化工具，平台技术，用户界面选项和高性能计算技

术等，项目旨在为医疗保险和医疗补助服务中心的重点项目使用行政索赔数据（医疗保险）来创建有用的信息产品以引导和支持决策。

食品与药物管理局（FDA）

虚拟实验室环境（VLE）将结合现有的资源和能力建立一个虚拟实验室数据网络，使用先进的分析和统计工具与功能，实现预测并促进公众健康、文档管理支持、电子临场能力等，促进世界范围内的合作，使任何地点在一小时内就如同一个虚拟实验室。

（本章编译者：周燕，清华大学国际传播研究中心助理研究员，硕士研究生）

参 考 文 献

[1] Adam Jacobs. 2009. The Pathologies of Big Data. Communications of the ACM. August (Vol. 52, No. 8): pp.36-44.

[2] Andrea Freyer Dugas, Yu-Hsiang Hsieh, Scott R. Levin, Jesse M. Pines, Darren P. Mareiniss, Amir Mohareb, Charlotte A. Gaydos, Trish M. Perl, and Richard E. Rothman. 2012. Google Flu Trends: Correlation With Emergency Department Influenza Rates and Crowding Metrics. Oxford Journals. Volume 54, Issue 4, pp.463-469.

[3] Begley S. 2011. The Best Medicine: The Quiet revolution in comparative effectiveness research just might save us from soaring medical costs, Scientific American 305, pp.50-55.

[4] Bella Hurrell, Andrew Leimdorfer. 2011. Data Journalism at the BBC. The Data Journalism Handbook. http://www.datajournalismhandbook.org/1.0/en/in_the_newsroom_1.html.

[5] Brian Boyer. 2011. How the News Apps Team at Chicago Tribune Works. The Data Journalism Handbook. http://datajournalismhandbook.org/1.0/en/in_the_newsroom_2.html.

[6] Brian G. Frizzelle, Kelly R. Evenson. 2009. The importance of accurate road data for spatial applications in public health: customizing a road network, International Journal of Health Geographics. http://www.datajournalismhandbook.org/1.0/en/in_the_newsroom_2.html.

[7] Brodie M.L., Greaves M., Hendler J.A. 2011. Databases and AI: The Twain Just Met, 2011 STI Semantic Summit, Riga, Latvia.

[8] Christian Bizer, Peter Boncz, Michael L. Brodie, Orri Erling. 2011. The Meaningful Use of Big Data: Four Perspectives - Four Challenges, SIGMOD Record, December (Vol. 40, No. 4), pp.56-60.

[9] Cooper James. 2009. Challenges for Database Management in the Internet of Things, IETE Technical Review. (Vol 26, ISSUE 5): SEP-OCT.

[10] Danah Boyd. 2010. Privacy and Publicity in the Context of Big Data. April 29. http://www.danah.org/papers/talks/2010/www2010.html.

[11] Danyel Fisher, Rob DeLine, Mary Czerwinski, Steven Drucker. 2012. Interactions with Big Data Analytics. Interactions, (Vol 3), pp.50-59.

[12] Dan Woods. 2011. Tableau Software's Pat Hanrahan on "What Is

a Data Scientist?” Forbes. Nov. 30. http://www.forbes.com/sites/danwoods/2011/11/30/tableau-softwares-pat-hanrahan-on-what-is-adata-scientist/2/.

[13] Doug Laney. 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Application Delivery Strategies, META Group.

[14] Edd Dumbill. 2012. Planning for Big Data, O'Reilly Media, Inc.

[15] Emily Bell. 2012. How a new research effort will help newsrooms determine what's next. April 30. http://www.knightfoundation.org/blogs/knightblog/2012/4/30/emily-bell-how-new-research-effort-will-help-newsrooms-determine-whats-next/.

[16] Encyclopedia Britannica. Data Mining. Encyclopedia Britannica Online Academic Edition. http://www.britannica.com/EBchecked/topic/1056150/data-mining.

[17] Erik Brynjolfsson, Lorin M. Hitt, Heekyung Hellen Kim. 2011. Strength in numbers: How does data-driven decision making affect firm performance? Social Science Research Network (SSRN), April.

[18] Executive Office of the President. 2012. Big Data Fact Sheet. March 29, http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf.

[19] Fauske, Kjell Magne. 2006. Neural Network. Digital image. Texample.net. 7 Dec. http://www.texample.net/tikz/examples/neural-network/.

[20] Frizzelle, Brian G, Evenson, Kelly R. 2009. The importance of accurate road data for spatial applications in public health: customizing a road network, International Journal of Health Geographics, 8:24.

[21] Gantz J, Reinsel D. 2011. Extracting Value from Chaos. Rep. IDC, Sponsored by EMC Corporation, June. http://idcdocserv.com/1142.

[22] Geoff McGhee. 2009. Journalism in the Age of Data. http://datajournalism.stanford.edu/noflash.html#.

[23] Gore A. I. 1998. The digital earth: understanding our planet in the 21st century. Given at the California Science Center. Los Angeles, January 31.

[24] Hardy Quentin. 2011. The Big Business of ‘Big Data’. October 24, http://bits.blogs.nytimes.com/2011/10/24/big-data/.

[25] Hurrell Bella, Leimdorfer Andrew: Data Journalism at the BBC. The Data Journalism Handbook. http://www.datajournalismhandbook.org/1.0/en/in_the_

newsroom_1.html.
[26] Ibarguen Bruce, Kar Kohinoor. 2009. Institute of Transportation Engineers. ITE Journal 79. 7 (Jul): pp. 30-32, 37-39.
[27] IBM: 智慧城市白皮书 . 2009. IBM. August.
[28] James X. Dempsey, Lara M. Flint. 2004. Commercial Data and National Security, George Washington Law Review, August, Vol. 72, No. 6.
[29] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. Annual Report: (6).
[30] Janna Quitney Anderson, Lee Rainie. 2012. Big Data: Experts say new forms of information analysis will help people be more nimble and adaptive, but worry over humans' capacity to understand and use these new tools well. Pew Research Center's Internet & American Life Project. July 20.
[31] Jeffrey F. Rayport. 2011. What Big Data Needs: A Code of Ethical Practices, Technology Review, May 26, http://www.technologyreview.com/business/37548/.
[32] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data, Nature, Vol. 457. Feb 19, pp. 1012-1015.
[33] Jonathan Gray, Liliana Bounegru, Lucy Chambers. 2011. The Data Journalism Handbook. http://datajournalismhandbook.org/1.0/en/.
[34] Joshua Cooper, Anne James. 2009. Challenges for Database Management in the Internet of Things, IETE Technical Review. (Vol. 26, ISSUE 5): SEP-OCT.
[35] Julia Angwin. 2010. The Web's New Gold Mine: Your Secrets, The Wall Street Journal, July 30, http://online.wsj.com/article/SB10001424052748703940904575395073512989404.html.
[36] Julia Angwin, Jennifer Valentino-Devries. 2011. Apple, Google Collect User Data, The Wall Street Journal, April 21, http://online.wsj.com/article/SB10001424052748703983704576277101723453610.html.
[37] Katherine Rowland. 2012. Epidemiologists put social media in the spotlight. Nature: 14 February.
[38] Mac Slocum. 2012. Big Data in Europe. April 23, http://radar.oreilly.com/2012/04/big-data-in-europe.html.

[39] Matthieu Pélissié du Rausas, James Manyika, Eric Hazan, Jacques Bughin, Michael Chui, Rémi Said. 2011. Internet matters: The Net's sweeping impact on growth, jobs, and prosperity. McKinsey Global Institute. Annual Report: (5).

[40] McKinsey Global Institute: Big data: The next frontier for innovation, competition, and productivity, Annual Report, 2011.6.

[41] Michael Chui, Markus Löffler, Roger Roberts. 2010. The Internet of Things. McKinsey Quarterly, March.

[42] Michael Stonebraker, Jason Hong. 2012. Researchers' Big Data Crisis: Understanding Design and Functionality, Communications of the ACM, Feb, Vol. 55, No. 2, pp.10-11.

[43] Mirko Lorenz. 2011. Why Journalists Should Use Data. The Data Journalism Handbook. http://www.datajournalismhandbook.org/1.0/en/introduction_1.html.

[44] Ovidiu Vermesan, Mark Harrison, Harald Vogt, Kostas Kalaboukas, Maurizio Tomasella, Karel Wouters, Sergio Gusmeroli, Stephan Haller. 2009. Internet of Things Strategic Research Roadmap, September 15.

[45] Paul Bradshaw. 2011. What Is Data Journalism? The Data Journalism Handbook. http://www.datajournalismhandbook.org/1.0/en/introduction_0.html.

[46] Pyle Dorian. 1999. Data Preparation for Data Mining. San Francisco, CA: Morgan Kaufmann.

[47] Quentin Hardy. 2011. The Big Business of 'Big Data'. October 24, http://bits.blogs.nytimes.com/2011/10/24/big-data/.

[48] Rebecca Hersher. 2012. Internet data miners' strike disease detection gold. Nature. Feb, (Vol.18, No.2): p.185.

[49] Simon Rogers. 2011. Behind the Scenes at the Guardian Datablog. The Data Journalism Handbook. http://www.datajournalismhandbook.org/1.0/en/in_the_newsroom_3.html.

[50] Sims D. 2012. The do's and don'ts of geo marketing. Commerce Weekly. http://radar.oreilly.com/2012/04/geofences-Google-wallet-facebook-mobile-carriers.html.

[51] Steve Lohr. 2012. The Age of Big Data, The New York Times, Februrary 11, http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all.

[52] Tom Heath, Christian Bizer. 2011. Linked Data: Evolving the Web Into a Global Data Space. Morgan & Claypool Publishers.
[53] Tomas Sanchez Lopez, Damith C. Ranasinghe, Mark Harrison, Duncan McFarlane. 2012. Adding sense to the Internet of Things: An architecture framework for Smart Object Systems. Pers Ubiquitous Computing (16), pp. 291-308.
[54] Vital Wave Consulting. 2012. Big Data, Big Impact: New Possibilities for International Development. The World Economic Forum Report.
[55] 曹磊，陈薇娜，缪其浩，陈超 . 2011. 大数据：数字世界的智慧基因 . 文汇报，12~21.
[56] 陈柳钦 . 2011. 智慧城市：全球城市发展新热点 . 青岛科技大学学报，(1).
[57] 陈效卫 . 2001. 防不胜防：信息时代国家安全面临的挑战 . 国际展望，(2).
[58] 戴维·奥尔森，石勇著 . 2007. 商业数据挖掘导论 . 吕巍等译 . 北京：机械工业出版社 .
[59] 甘绮翠，J. Chris Harreld, 姜一炜，李宇恒，赵建军 . 2009. 智慧地球赢在中国 . IBM 商业价值研究院 .
[60] 工业和信息化软件与集成电路促进中心 . 2009. 智慧地球的认识和思考，(09).
[61] 韩慧，毛峰，王文渊 . 2004. 数据挖掘中决策树算法的最新进展 . 计算机应用研究，(12).
[62] 刘强 . 2004. 网络恐怖主义的特性、现状及发展趋势 . 世界经济与政治论坛，(4).
[63] 李德仁，龚健雅，邵振峰 . 2010. 从数字地球到智慧地球 . 武汉大学学报，(2).
[64] 彭明盛 . 2008. 智慧地球下一代领导人议程 . November, http://www.ibtimes.com.cn/articles/20090224/zhihuidiqiu_3.htm.
[65] 秦洪花，李汉清，赵霞 . 2010. “智慧城市”的国内外发展现状 . 信息化建设，(9).
[66] 王杰聪 . 2012. 英特尔和中科院联合成立物联技术研究院 . 2012-04-11. http://tech.163.com/12/0411/14/7UQNAJJ600094MLL.html.
[67] 乌尔里希·贝克著 . 2004. 风险社会 . 何博闻译 . 南京：译林出版社 .
[68] 乌尔里希·贝克著 . 2003. 从工业社会到风险社会——关于人类生存、社会结构和生态启蒙等问题的思考 . 王武龙编译 . 马克思主义与现实，(3).
[69] 杨辅祥，刘云超，段智华 . 2002. 数据清理综述 . 计算机应用研究，(3).

[70] 叶涛 . 2012. 从蛛丝马迹中探寻事实真相——利用日志信息调查互联网用户的行为规律 . 调研世界 , (2).
[71] 俞晓秋 . 2002. 全球信息网络安全动向与特点 . 现代国际关系 , (2).
[72] 张旭 , 杜红亮 , 陈颖健 . 2011. 关于“智慧地球”战略的再认识及再思考 . 科学管理研究 , (17).
[73] 张永民 . 2010. 解读智慧地球与智慧城市 . 中国信息界 , (10).
[74] 郑慧会 , 李兴保 , 刘建美 . 2009. Web 3.0——网上学习新平台 . 现代教育技术 , (4).
[75] 中国电信智慧城市研究组 . 2011. 智慧城市之路 . 电子工业出版社 , (9).
[76] 中国三星经济研究院 . 2011. 全球智能城市发展模式比较 , (1).
[77] 周珍妮 , 陈碧荣 . 2009. Web 3.0——全新的互联网时代 . 图书情报论坛 , (1).

大数据相关术语[①]

（注：按照英文首字母排序）

A/B 测试（A/B testing）：在一个控制组和各个对照组比较中，发现哪种方案更有利于给定客观变量（如市场反应速度）发挥作用。这种测试也称为“对比测试”或“水桶测试”。这种测试可以应用在什么样的文本结构、图片、网页颜色可以提高一个商业网站的交易量。大数据能使巨大数量的测试具有可操作性和分析性。确保在足够规模的基础上探测控制组和试验组之间有统计学意义的差异。测试中出现多个变量的统计建模技术被称为 A/B/N 测试。

关联规则研究（Association rule learning）：数据挖掘中用来发现大数据库中各种变量之间有趣关系的一系列方法。应用之一就是“购物篮分析”（Market basket analysis），通过数据分析可以在门店的销售过程中找到具有关联关系的商品，并以此获得销售收益的增长。比如研究发现超市中购买啤酒的消费者也易于购买纸尿裤。

分类（Classification）：应用于数据挖掘的一项技术，指在训练数据集里已经分类的数据基础上归类新的数据，有异于聚类分析（Cluster analysis），被称为监督分析（supervised learning）。应用之一就是预测特定群体消费行为（购买决定、顾客流失率、消费率等）。

聚类分析（Cluster analysis）：一种物以类聚的多元统计方法，在预先对事物没有分类经验，即分类界面不清楚的情况下，可用聚类分析进行分类，并结合判别分析，对新事物作类别预测。应用之一为将消费者细分为针对不同目标市场的相似群体。

众包（Crowdsourcing）：指一些没有清晰界限的雇员、项目发起者或总包商、外包等组成的群体在执行传统任务中形成的一种行为。这是一种大规模合作和使用 Web2.0 的实例。

数据融合和数据集成（Data fusion and data integration）：集成和分析多源数据的一系列方法，旨在发展比单一来源数据的分析更有效和更具准确性的方式。信号处理技术可以用来实现某些类型的数据融合。此项技术的运用包括使用物联网中的传感器数据监测炼油厂的复杂分布系统表现。

数据挖掘（Data mining）：集合统计学、机械学及数据库管理的方法处理大型数据库的技术。这些方法包括关联规则研究、聚类分析、回归分析和归

① 资料翻译自：The McKinsey Global Institute, Big data: The next frontier for innovation, competition,and productivity, 2011.6。

类等。

系综研究（Ensemble learning）：指使用一系列学习器进行学习，并使用某种规则把各个学习结果进行整合从而获得比单个学习器更好的学习效果的一种机器学习方法。

基因演算法（Genetic algorithms）：以自然界生物基因中 DNA 编码与繁殖的原理，用以模拟自然环境与人造环境中的一些现象的研究方法。无论自然或人造环境，都可以将事物依其属性进行如基因 DNA 一样的编码，并在物群之间借由编码的运算繁衍出“下一代”。通过函数设计可以遴选适合环境的“下一代”继续参与繁衍，由此获得较适合环境的物种。基因演算法经常被描述为一种“进化算法”，适合解决非线性问题。

机器学习（Machine learning）：亦称“人工智能”，研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

自然语言的处理方法（Natural language processing）：是关于计算机与人类（自然语言）交互的计算机科学与语言学的一个领域。自然语言生成（Natural language generation）系统把计算机数据库信息转换成人类可读的语言。自然语言理解系统（Natural language understanding）把人类语言的样本转换为计算机容易处理的更加形式化的表示，诸如分析树或者一阶逻辑。NLP 范围内的很多难题适用于生成和理解，例如，为了理解句子，计算机必须能够为形态（morphology，词的构造）建模，为了生产语法上正确的英语句子也必须有形态的模型。

类神经网路（Neural networks）：以电脑（软件或硬件）来模拟生物大脑神经的人工智能系统，并将此应用于辨识、决策、控制、预测等行为的计算机模型。

网络分析（Network analysis）：在一个图表或网络中用来描述节点之间关系的一系列方法，在社会网络分析中，还可以是社区或组织中的个人链接分析。例如，信息是如何传播的，或者是谁的影响力最大等。典型的应用是分析市场营销中的主要意见领袖或确认企业信息流的瓶颈等。

最优化（Optimization）：指运用数学方法研究各种系统（如成本、速度、可靠性）的优化途径及方案，为决策者提供科学决策的依据。最优化方法的主要研究对象是各种有组织系统的管理问题及其生产经营活动。

模式识别（Pattern recognition）：指对表征事物或现象的各种形式的（数值的、文字的和逻辑关系的）信息进行处理和分析，以对事物或现象进行描述、辨认、分类和解释的过程，是信息科学和人工智能的重要组成部分。

预测建模（Predictive modeling）：建立和选择一个数学模型，用来最好预

测某种可能性的一系列方法。典型应用为顾客关系管理，预测顾客选择其他卖家或带来更多买家的可能性。

回归（Regression）：确定两种或两种以上变数间相互依赖的定量关系的一种统计分析方法，常用于预测和预报。如基于不同市场和经济变量的销售预期，或影响顾客满意度的生产参数。

情感分析（Sentiment analysis）：一种应用自然语言的处理和其他分析技术，从源文本文件中识别和提取信息的方法。主要为判定作者或者演讲者对某个话题的态度，包括判断、评价、情绪状况、情绪交流等。企业使用情感分析方法，可以通过社交媒体（博客、微博、社会网络）发现细分的客户群及股民对公司产品和行为的反应等。

信号处理（Signal processing）：一种来自于电气工程和应用数学的方法，最初用以分析离散和连续信号，如无线电信号、声音和图像等。

空间分析（Spatial analysis）：空间分析是基于地理对象的位置和形态的空间数据的分析技术，其目的在于提取和传输空间信息。 如关联性分析（分析顾客购买商品的意愿与商品摆放位置有无关系）或模拟（同一企业在不同地区生产供应链的表现）。

统计学（Statistics）：收集、组织和解释调查设计及实验数据的科学。统计技术经常被用来判断变量之间关系的偶然性（无效假设），以及变量之间可能存在的因果关系（显著性检验）。如 A/B 测试即是为了检验什么样的市场因素最能增加收入。

监督学习（Supervised learning）：一种机器学习技术，可以由训练资料中学到或建立一个模式（Learning model），并依此模式推测新的模型。训练资料是由输入物件（通常是向量）和预期输出所组成。函数的输出可以是一个连续的值（称为回归分析）。或是预测一个分类标签（称作分类）。与无监督学习不同。

模拟（Simulation）：复杂系统的行为建模，通常用于预测、估算和情景规划。例如蒙特卡洛模拟，是一类通过设定随机过程，反复生成时间序列，计算参数估计量和统计量，进而研究其分布特征的方法。其结果是给出了一个概率分布的直方图。比如评估可能满足财务目标的各种举措的成功可能性。

时间序列分析（Time series analysis）：一种动态数据处理的统计方法。该方法基于随机过程理论和数理统计学方法，研究随机数据序列所遵从的统计规律，以用于解决实际问题。时间序列分析的例子包括股市指数的时值或某种给定条件下每天确诊患者的人数。

无监督学习（Unsupervised learning）：发现隐藏在未标记的数据结构中的

一套机器学习技术。聚类分析是无指导学习的范例。

可视化（Visualization）：创建图片、图表、动画，用以沟通、理解、提高大数据分析结果的技术。

大数据技术（Big data technologies）：用于聚合、操作、管理和分析大数据的技术。

大表格（Big table）：建立在谷歌文件系统上的专有分布式数据库系统。

商业智能（Business intelligence）：用于设计报告、分析和呈现数据的应用软件。BL 工具往往用于读取已存储在数据仓库或数据集市中的数据，也可以用来定期生成标准报告，或在实时管理仪表盘显示信息。

卡桑德拉（Cassandra）：一个免费的开源数据库管理系统，用以处理大量的分布式系统中的数据。该系统最初是在脸谱上开发，现作为 Apache 软件基金会的一个项目运行。

云计算（Cloud computing）：基于互联网的相关服务的增加、使用和交付模式，通常涉及通过互联网来提供动态易扩展且经常是虚拟化的资源。云是网络、互联网的一种比喻说法。过去在图中往往用云来表示电信网，后来也用来表示互联网和底层基础设施的抽象。

数据集市（Data mart）：数据仓库的子集，通常通过商业智能工具向用户提供数据。

数据仓库（Data warehouse）：优化报告的专门数据库，通常用于存储大量的结构化数据。数据使用 ETL（提取、转换和加载）工具从业务数据存储中上传，使用商业智能工具生成报告。

分布式系统（Distributed system）：多台计算机通过网络通信，用以解决常见的计算问题。问题分为多个任务，每个由一台或多台并行工作的计算机解决。分布式系统的优点包括更高的性能，较低的成本，更高的可靠性，以及更多的可扩展性。

发电机（Dynamo）：专有分布式数据存储系统，由亚马逊开发。

提取、转换和加载（Extract, Transform and Load，ETL）：一种软件工具，用于从外部资源中提取数据、转换以适应操作需求，并将其装入一个数据库或数据仓库。

谷歌文件系统（Google file system）：专有分布式文件系统，由谷歌开发。

平台（Hadoop）：一个分布式系统基础架构。用户可以在不了解分布式底层细节的情况下，开发分布式程序。充分利用集群的威力高速运算和存储，Hadoop 实现了一个分布式文件系统（Hadoop Distributed File System），简称 HDFS。

HBase（HBase）：一个免费的开源、分布式、非关系型数据库，以谷歌的大表格为蓝本。

映射化简算法（MapReduce）：谷歌开发的 C++ 编程工具，用于大规模数据集（大于 1TB）的并行运算。

混搭（Mashup）：一种应用程序，使用并结合来自两个或多个数据源的数据演示或功能，以创造新的服务。这些应用往往在网络上，通过开放应用编程接口或开放的数据源访问使用数据。

元数据（Metadata）：描述数据文件的内容和背景的数据。

非关系型数据库（Non-relational database）：不将数据存在表中的数据库，与结构型数据库相对。

R 语言（R）：用于统计分析、绘图的语言和操作环境，属于 GNU 系统的一个自由、免费、源代码开放的软件，是用于统计计算和统计制图的优秀工具。

关系型数据库（Relational database）：是建立在关系模型基础上的数据库，借助于集合代数等数学概念和方法来处理数据库中的数据。 标准数据查询语言 SQL 就是一种基于关系数据库的语言，这种语言执行对关系数据库中数据的检索和操作。关系模型由关系数据结构、关系操作集合、关系完整性约束三部分组成。

半结构数据（Semi-structured data）：字段数目不定的数据，如 Exchange 存储的数据。

结构化查询语言（Structured Query Language）：一种数据库查询和程序设计语言，用于存取数据以及查询、更新和管理关系数据库系统，SQL 也是数据库脚本文件的扩展名。

流处理（Stream processing）：也称事件流处理，指处理大量实时事件数据流的技术设计。流处理使金融服务中的交易算法、RFID 事件处理应用程序、欺诈检测、过程监控、基于位置电信服务成为可能。

结构化数据（Structured data）：以二维表结构存储在数据库中的数据，如常用的 Excel 软件所处理的数据。

非结构化数据（Unstructured data）：不能用数据库二维逻辑表来表现的数据即称为非结构化数据，包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频 / 视频信息等。

可视化（Visualization）：用于创建图片、图表或动画的技术，经常被用来合成大数据分析的结果。

（附录部分编译者为刘娟，清华大学国际传播研究中心助理研究员，博士研究生）